Advances in Intelligent Systems and Computing 841

Vijay Singh Rathore · Marcel Worring
Durgesh Kumar Mishra · Amit Joshi
Shikha Maheshwari *Editors*

Emerging Trends in Expert Applications and Security

Proceedings of ICETEAS 2018

Springer

Advances in Intelligent Systems and Computing

Volume 841

Series editor

Janusz Kacprzyk, Polish Academy of Sciences, Warsaw, Poland
e-mail: kacprzyk@ibspan.waw.pl

drksujatha23@gmail.com

The series "Advances in Intelligent Systems and Computing" contains publications on theory, applications, and design methods of Intelligent Systems and Intelligent Computing. Virtually all disciplines such as engineering, natural sciences, computer and information science, ICT, economics, business, e-commerce, environment, healthcare, life science are covered. The list of topics spans all the areas of modern intelligent systems and computing such as: computational intelligence, soft computing including neural networks, fuzzy systems, evolutionary computing and the fusion of these paradigms, social intelligence, ambient intelligence, computational neuroscience, artificial life, virtual worlds and society, cognitive science and systems, Perception and Vision, DNA and immune based systems, self-organizing and adaptive systems, e-Learning and teaching, human-centered and human-centric computing, recommender systems, intelligent control, robotics and mechatronics including human-machine teaming, knowledge-based paradigms, learning paradigms, machine ethics, intelligent data analysis, knowledge management, intelligent agents, intelligent decision making and support, intelligent network security, trust management, interactive entertainment, Web intelligence and multimedia.

The publications within "Advances in Intelligent Systems and Computing" are primarily proceedings of important conferences, symposia and congresses. They cover significant recent developments in the field, both of a foundational and applicable character. An important characteristic feature of the series is the short publication time and world-wide distribution. This permits a rapid and broad dissemination of research results.

Advisory Board

Chairman

Nikhil R. Pal, Indian Statistical Institute, Kolkata, India
e-mail: nikhil@isical.ac.in

Members

Rafael Bello Perez, Universidad Central "Marta Abreu" de Las Villas, Santa Clara, Cuba
e-mail: rbellop@uclv.edu.cu

Emilio S. Corchado, University of Salamanca, Salamanca, Spain
e-mail: escorchado@usal.es

Hani Hagras, University of Essex, Colchester, UK
e-mail: hani@essex.ac.uk

László T. Kóczy, Széchenyi István University, Győr, Hungary
e-mail: koczy@sze.hu

Vladik Kreinovich, University of Texas at El Paso, El Paso, USA
e-mail: vladik@utep.edu

Chin-Teng Lin, National Chiao Tung University, Hsinchu, Taiwan
e-mail: ctlin@mail.nctu.edu.tw

Jie Lu, University of Technology, Sydney, Australia
e-mail: Jie.Lu@uts.edu.au

Patricia Melin, Tijuana Institute of Technology, Tijuana, Mexico
e-mail: epmelin@hafsamx.org

Nadia Nedjah, State University of Rio de Janeiro, Rio de Janeiro, Brazil
e-mail: nadia@eng.uerj.br

Ngoc Thanh Nguyen, Wroclaw University of Technology, Wroclaw, Poland
e-mail: Ngoc-Thanh.Nguyen@pwr.edu.pl

Jun Wang, The Chinese University of Hong Kong, Shatin, Hong Kong
e-mail: jwang@mae.cuhk.edu.hk

More information about this series at http://www.springer.com/series/11156

drksujatha23@gmail.com

Vijay Singh Rathore · Marcel Worring
Durgesh Kumar Mishra · Amit Joshi
Shikha Maheshwari
Editors

Emerging Trends in Expert Applications and Security

Proceedings of ICETEAS 2018

Springer

drksujatha23@gmail.com

Editors
Vijay Singh Rathore
Jaipur Engineering College
and Research Centre
Jaipur, Rajasthan, India

Marcel Worring
Intelligent Systems Lab
University of Amsterdam
Amsterdam, The Netherlands

Durgesh Kumar Mishra
Sri Aurobindo Institute of Technology
Indore, Madhya Pradesh, India

Amit Joshi
Sabar Institute of Technology for Girls
Ahmedabad, Gujarat, India

Shikha Maheshwari
Department of Computer Science
and Engineering
Jaipur Engineering College
and Research Centre
Jaipur, Rajasthan, India

ISSN 2194-5357 ISSN 2194-5365 (electronic)
Advances in Intelligent Systems and Computing
ISBN 978-981-13-2284-6 ISBN 978-981-13-2285-3 (eBook)
https://doi.org/10.1007/978-981-13-2285-3

Library of Congress Control Number: 2018951909

© Springer Nature Singapore Pte Ltd. 2019
This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

drksujatha23@gmail.com

Preface

The *International Conference on Emerging Trends in Expert Applications and Security* (ICETEAS 2018) has been held at Jaipur, India, during February 17–18, 2018. ICETEAS 2018 has been organized and supported by the JECRC Foundation, Jaipur, India. The conference addressed innovative advancements in expert applications and security issues therein with specific reference to the area of computing.

Nowadays, globalization of academic and applied research is growing at great pace. Computer, communication, and computational sciences are hot areas with a lot of thrust. Keeping this ideology in preference, JECRC Foundation, Jaipur, India, has come up with the international event. ICETEAS 2018 is an international opportunistic forum and vibrant platform for researchers and industry practitioners to exchange of state-of-the-art knowledge gained from their original research work and practical development.

The expert application concept of the conference focused on current advances in the research and use of the expert system with particular focus on the role in maintaining academic level while sharing novel research and cutting-edge developments in the fields of computer system security using cryptographic algorithms and other security schemes for the system as well as the cloud. The outcome of the conference significantly contributes to the society by serving the expert community seeking to stimulate the development to improve lives throughout the world by providing innovative results.

ICETEAS 2018 has a foreseen objective of enhancing the research activities at a large scale. ICETEAS 2018, first of all, is an abbreviation of the full name of the conference, i.e., *International Conference on Emerging Trends in Expert Applications and Security*. Another meaning breaks it into ICE and TEAS, wherein ICE represents the hurdles, challenges, problems, and limitations being faced by the researchers in the pursuance of their research practices and implementation of innovations and TEAS is the knowledge dissemination and exchange of ideas to melt ICE. Through ICETEAS 2018, all the concerned people will be provided a genuine platform with all possible solutions and opportunities to find ways to come out of their research problems and limitations.

drksujatha23@gmail.com

One more idea to keep this name ICETEAS 2018 is that this conference is conducted mostly in the winter season (here ICE denotes Cold), and this Cold ICE is melted during the various TEA sessions and TEA breaks (TEAS) wherein the participants find many opportunities to collect knowledge, explore more, and establish strong networking and platform for sharing each other's ICE melting ideas. In true sense, ICETEAS 2018 is an international ICE breaking and melting platform through TEAS of great knowledge, innovations, and networking opportunities. Technical Program Committee and Advisory Board of ICETEAS include eminent academicians, researchers, and practitioners from abroad as well as from all over the nation.

ICETEAS 2018 received around 367 submissions from around 639 authors of 10 different countries such as UK, USA, Netherlands, Italy, Japan, China, Korea, Thailand, Vietnam, Taiwan. Each submission has gone through the plagiarism check. On the basis of plagiarism report, each submission was rigorously reviewed by at least three reviewers with an average of 2.07 per reviewer. Even some submissions have more than three reviews. On the basis of these reviews and presentation during the conference, 86 high-quality papers were selected for publication in this proceedings volume, with an acceptance rate of 23%.

We are thankful to the speakers, delegates, and the authors for their participation and their interest in ICETEAS 2018 as a platform to share their ideas and innovation. We are also thankful to Mr. Aninda Bose, Senior Editor, Hard Sciences, Springer, for providing continuous guidance and support. Also, we extend our heartfelt gratitude and thanks to the Reviewers and Technical Program Committee Members for showing their concern and efforts in the review process. We are indeed thankful to everyone directly or indirectly associated with the conference organizing a team, leading it toward the success.

We hope you enjoy the conference proceedings and wish you all the best.

Jaipur, India

Vijay Singh Rathore
Marcel Worring
Durgesh Kumar Mishra
Amit Joshi
Shikha Maheshwari
ICETEAS 2018

drksujatha23@gmail.com

Acknowledgements

The organization of a conference is always a stressful adventure because of all the very small things and all the very important issues that have to be planned and managed.

We are grateful for the extensive support of Shri O. P. Agrawal, Chief Patron; Shri Amit Agrawal, Patron and Director, JECRC Foundation; and Shri Arpit Agrawal, Director, JECRC, Jaipur, during the planning as well as implementation phase of this conference to make this conference possible. We take the opportunity to thank our Principal, Dr. Vinay Kumar Chandna, for his very friendly support and advice since the conception of this idea.

We are grateful to a number of people without whom we would not have been able to successfully organize this mega event and personally thankful to Prof. Marcel Worring, Director, Informatics Institute, University of Amsterdam, the Netherlands; Prof. Sheng Lung Peng, Professor, Computer Science and Information Engineering, National Dong Hwa University, Hualien, Taiwan; Dr. Kirti Seth, Associate Professor, SOCIE, Inha University, Tashkent Uzbekistan; Prof. Vincenzo Piuri, University of Milan, Italy; Prof. Rajeev Gupta, Pro Vice-Chancellor, Rajasthan Technical University, Kota; Prof. C. P. Gupta, Rajasthan Technical University, Kota; Dr. O. P. Rishi, Associate Professor, University of Kota, Kota; Dr. Durgesh Kumar Mishra, Chairman, CSI-(Div IV); Prof. Peter H. Kent, CEO, UKEI Ltd., London; Prof. David Wing, Director, UKEI Ltd., London; Prof. M. Hanumanthappa, Director, Department of Computer Science, Bangalore University, Bangalore; Prof. P. V. Virparia, Sardar Patel University, Gujarat; Prof. Atul Gonsai, Saurashtra University, Rajkot, India; Prof. C. K. Kumbharana, Saurashtra University, Rajkot; Prof. Bankim Patel, UT University, Baroda; Prof. Vibhakar Mansotra, Professor, University of Jammu; Dr. Pawanesh Abrol, Professor, University of Jammu, Jammu; Dr. Vinod Sharma, Professor, University of Jammu, Jammu; Dr. Jatinder Singh Manhas, Associate Professor, University of Jammu, Jammu; Dr. Vivek Tiwari, Assistant Professor; Dr. Shyama Prasad Mukherjee International Institute of Information Technology, Chhattisgarh; Dr. Paras Kothari, Principal, Samarth Group of Institutions, Himmat Nagar, Gujarat; Dr. K. Baskaran, Associate Professor and

drksujatha23@gmail.com

Head of the Department, Department of Electronics and Instrumentation Engineering, Government College of Technology Coimbatore; Prof. P. K. Mishra, Professor, Banaras Hindu University; Dr. Tanupriya Choudhury, Associate Professor, Amity University, Noida; Dr. Praveen Kumar, Associate Professor, Amity School of Engineering and Technology, Amity University, Noida; Dr. Sumeet Gill, Associate Professor, Maharshi Dayanand University, Rohtak; Dr. S. S. Dalal, Associate Professor, SRM University, Haryana; Dr. Bhavna Arora, Assistant Professor, Central University, Jammu; Dr. Kusum Rajawat, Principal, Shree Karni College, Jaipur; Dr. M. Venkatesh Kumar, Chairman, IEEE Young Professionals, Tamil Nadu Circle; Dr. Bharat Singh Deora, Associate Professor, Computer Science, JRN Rajasthan Vidyapeeth University, Udaipur; Dr. M. N. Hoda, Director, BVICAM, New Delhi; Prof. Vipin Tyagi, JUET, India, and Vice President, Region 3, CSI; Dr. Nilanjan Dey, Editor, IGI Global Journal, and all other colleagues of the state of Rajasthan.

We would like to thank our esteemed authors for having shown confidence in us and considered ICETEAS 2018 a platform to showcase and share their original research work and came to Jaipur to present it as well.

We wish to express our sincere gratitude to the focused team of Chairs, Co-chairs, International Advisory Committee, and Technical Program Committee.

We will fail in our duty, if we do not thank our Publishing Co-partner Shri Amit Joshi, CEO, and Shri Mihir Chauhan, Director, GR Foundation, who worked constantly behind the scene and giving us all encouragement, support, and a path aimed to quality and excellence, comparable to the best in the world. We are also thankful to Shri Aninda Bose, Senior Editor, Hard Sciences, Springer, for providing continuous guidance and support.

Our heartfelt appreciation to Ms. Shikha Maheshwari for coming forward to undertake the challenge so actively, taking ownership and together making this conference a great success.

We would like to address a particular warm thank to the colleagues of the Department of Computer Science and Engineering, JECRC, Jaipur, for their participation and expertise in the preparation of the conference. We deeply acknowledge the support provided by the other department colleagues for management of conference properly.

Finally, we are thankful to one and all, who have contributed directly or indirectly in making this conference ICETEAS 2018 successful.

Thanks to Almighty for everything.

Jaipur, India
March 2018

Vijay Singh Rathore
Marcel Worring
Durgesh Kumar Mishra
Amit Joshi
Shikha Maheshwari

drksujatha23@gmail.com

About This Book

This book includes high-quality and peer-reviewed papers from the *International Conference on Emerging Trends in Expert Applications and Security* (ICETEAS 2018), held at Jaipur Engineering College and Research Centre, Jaipur, India, on February 17–18, 2018, presenting the latest developments and technical solutions in expert applications and security. Expert applications and security is receiving increasing popularity and acceptance in the engineering community because of the existence of a close match between the capabilities of the current generation expert systems and the requirements of engineering practice. Keeping this ideology in mind, the book offers insights that reflect the advances in these fields from upcoming researchers and leading academicians across the globe. Covering a variety of topics, such as expert applications and artificial intelligence, information and application security, advanced computing, multimedia applications in forensics, security and intelligence, advances in web technologies, and implementation and security issues, it helps those in the computer industry and academia use the advances of next-generation communication and computational technology to shape real-world applications. The book is appropriate for the researcher and the professional. The researcher can save considerable time in searching the scattered technical information on expert applications and security. The professional can have a readily available rich set of guidelines and techniques that are applicable to a wide class of engineering domains.

drksujatha23@gmail.com

Contents

Nuts and Bolts of ETL in Data Warehouse 1
Sharma Sachin, Sandip Kumar Goyal, Sharma Avinash and Kumar Kamal

User Identification Over Digital Social Network Using Fingerprint Authentication 11
Devender Dhaked, Surendra Yadav, Manish Mathuria and Saroj Agrawal

A Review on Machine Translation Systems in India 23
Shikha Maheshwari, Prashant S. Saxena and Vijay Singh Rathore

Fuzzy-Based Analysis of Information Security Situation 31
Ashish Srivastava and Pallavi Shrivastava

Estimation of Microwave Dielectric Constant Using Artificial Neural Networks 41
K. Sujatha, R. S. Ponmagal, G. Saravanan and Nallamilli P. G. Bhavani

Bone Fracture Detection from X-Ray Image of Human Fingers Using Image Processing 47
Anil K. Bharodiya and Atul M. Gonsai

Review of Data Analysis Framework for Variety of Big Data 55
Yojna Arora and Dinesh Goyal

Time Series Forecasting of Gold Prices 63
Saim Khan and Shweta Bhardwaj

Titanic Data Analysis by R Data Language for Insights and Correlation 73
Shaurya Khanna, Shweta Bhardwaj and Anirudh Khurana

Edge Detection Property of 2D Cellular Automata 81
Wani Shah Jahan

drksujatha23@gmail.com

Augmented Intelligence: A Way for Helping Universities to Make Smarter Decisions 89
Manu Sharma

A Multiple String and Pattern Matching Algorithm Using Context-Free Grammar 97
Sarvesh Kumar, Sonali Singh, Arfiha Khatoon and Swati Agarwal

A Review of Machine Translation Systems for Indian Languages and Their Approaches 103
Dipal Padhya and Jikitsha Sheth

Energy-Efficient Cloud Computing for Smart Phones 111
Nancy Arya, Sunita Chaudhary and S. Taruna

A Bounding Box Approach for Performing Dynamic Optical Character Recognition in MATLAB 117
Poonam Chaturvedi, Mohit Saxena and Bhavna Sharma

Performance Comparison of LANMAR and AODV in Heterogenous Wireless Ad-hoc Network 125
Madhavi Dhingra, S. C. Jain and Rakesh Singh Jadon

An Efficient Approach for Power Aware Routing Protocol for MANETs Using Genetic Algorithm 133
Renu Choudhary and Pankaj Kumar Sharma

Multi-purposed Question Answer Generator with Natural Language Processing 139
Hiral Desai, Mohammed Firdos Alam Sheikh and Satyendra K. Sharma

Building Machine Learning Based Diseases Diagnosis System Considering Various Features of Datasets 147
Shrwan Ram and Shloak Gupta

Enhancing Data Security in Cloud Using Split Algorithm, Caesar Cipher, and Vigenere Cipher, Homomorphism Encryption Scheme 157
Abhishek Singh and Shilpi Sharma

k-dLst Tree: k-d Tree with Linked List to Handle Duplicate Keys 167
Meenakshi and Sumeet Gill

Feature Extraction in Geospatio-temporal Satellite Data for Vegetation Monitoring 177
Hemlata Goyal, Nisheeth Joshi and Chilka Sharma

Multiple Objects Tracking Under Occlusion Detection in Video Sequences 189
Sanjay Gaur, Sheshang Degadwala and Arpana Mahajan

drksujatha23@gmail.com

IoT Platform for Smart City: A Global Survey 197
Rakesh Roshan, Anukrati Sharma and O. P. Rishi

Incessant Ridge Estimation Using RBCA Model 203
Sandeep Kumar Sharma, C. S. Lamba and Vijay Singh Rathore

Impact of Try A-Gain—An Online Game App for Society 211
Vijay Singh Rathore, Shikha Maheshwari, Diwanshu Soni, Apoorva Agrawal, Ayush Khandelwal, Aman Vijay, Divyang Bhargava and Aniket Dixit

A Comparative Analysis of Wavelet Families for Invisible Image Embedding 219
Neha Solanki, Sarika Khandelwal, Sanjay Gaur and Diwakar Gautam

A Video Database for Intelligent Video Authentication 229
Priya Gupta, Ankur Raj, Shashikant Singh and Seema Yadav

Software Quality Improvement Through Penetration Testing 239
Subhash Chandra Jat, C. S. Lamba and Vijay Singh Rathore

Air Pollution Concentration Calculation and Prediction 245
Jyoti Gautam, Arushi Gupta, Kavya Gupta and Mahima Tiwari

The NLP Techniques for Automatic Multi-article News Summarization Based on Abstract Meaning Representation 253
Deepa Nagalavi and M. Hanumanthappa

An Analysis of Load Management System by Using Unified Power Quality Conditioner for Distribution Network 261
D. Jayalakshmi, S. Sankar and M. Venkateshkumar

Design and Comparative Analysis of Various Intelligent Controller Based Efficiency Improvement of Fuel Cell System 273
M. Venkateshkumar, R. Raghavan, R. Indumathi and Shivashankar Sukumar

Analysis of Load Balancing Algorithms Using Cloud Analyst 291
Jyoti Rathore, Bright Keswani and Vijay Singh Rathore

Terrain Attribute Prediction Modelling for Southern Gujarat: A Geo-spatial Perspective 299
Jaishree Tailor and Kalpesh Lad

Sentiment Analysis of Live Tweets After Elections 307
Palak Baid and Neelam Chaplot

Smart Innovation Regarding Bringing Kitchen Food Items in the Kitchen by Automatically Informing the Shopkeeper by Using GSM 900 Board and Arduino Uno R3 Board with Proper Programming 315
Vijay Kumar, Vipul Sharma and Avinash Sharma

drksujatha23@gmail.com

Sentence Tokenization Using Statistical Unsupervised Machine Learning and Rule-Based Approach for Running Text in Gujarati Language 319
Chetana Tailor and Bankim Patel

A Hybrid Approach to Authentication of Signature Using DTSVM 327
Upasna Jindal and Surjeet Dalal

Securing Web Access—DNS Threats and Remedies 337
Anchal Sehgal and Abhishek Dixit

P2S_DLB: Pluggable to Scheduler Dynamic Load Balancing Algorithm for Distributed Computing Environment 347
Devendra Thakor and Bankim Patel

Parkinson Disease Prediction Using Machine Learning Algorithm 357
Richa Mathur, Vibhakar Pathak and Devesh Bandil

Hybrid Technique Based on DBSCAN for Selection of Improved Features for Intrusion Detection System 365
Akash Saxena, Khushboo Saxena and Jayanti Goyal

A Study on Performance Evaluation of Cryptographic Algorithm 379
Mohammed Firdos Alam Sheikh, Sanjay Gaur, Hiral Desai and S. K. Sharma

Optimal Ant and Join Cardinality for Distributed Query Optimization Using Ant Colony Optimization Algorithm 385
Preeti Tiwari and Swati V. Chande

Comparative Study of Various Cryptographic Algorithms Used for Text, Image, and Video 393
Nilesh Advani, Chetan Rathod and Atul M. Gonsai

A Comparative Study of Ontology Building Tools for Contextual Information Retrieval 401
Ripal Ranpara, Asifkhan Yusufzai and C. K. Kumbharana

A Comparative Study of Cryptographic Algorithms for Cloud Security 409
Asifkhan Yusufzai, Ripal Ranpara, Mital Vora and C. K. Kumbharana

Mathematical Modelling and Analysis of Graphene Using Simulink Technique 417
Pragati Tripathi and Shabana Urooj

Development of IoT for Smart Agriculture a Review 425
Kamlesh Lakhwani, Hemant Gianey, Niket Agarwal and Shashank Gupta

drksujatha23@gmail.com

Multi-label Classification Trending Challenges and Approaches 433
Pooja Pant, A. Sai Sabitha, Tanupriya Choudhury and Prince Dhingra

Virtual Reality as a Marketing Tool 445
Harry Singh, Chetna Singh and Rana Majumdar

A Critical and Statistical Analysis of Air Pollution Using Data Analytics 451
Praveen Kumar, Paras Lalwani, Karan Rathore and Seema Rawat

Conceptual Structure of ASMAN Framework to Compare SaaS Providers 461
Mamta Dadhich and Vijay Singh Rathore

A Case Study of Feedback as Website Design Issue 469
Jatinder Manhas, Amit Sharma, Shallu Kotwal and Viverdhana Sharma

A Review of Sentimental Analysis on Social Media Application 477
Akankasha and Bhavna Arora

Trust Prediction Using Ant Colony Optimization and Particle Swarm Optimization in Social Networks 485
Rajeev Goyal, Arvind K. Updhyay and Sanjiv Sharma

Stock Market Price Prediction Using LSTM RNN 493
Kriti Pawar, Raj Srujan Jalem and Vivek Tiwari

Honeypots and Its Deployment: A Review 505
Neeraj Bhagat and Bhavna Arora

Study on Data Mining with Drug Discovery 513
Bahul Diwan and Shweta Bhardwaj

Efficient Hybrid Recommendation Model Based on Content and Collaborative Filtering Approach 521
Ankita Gupta, Alok Barddhan, Nidhi Jain and Praveen Kumar

Research Review on Digital Image Steganography Which Resists Against Compression 529
Darshan M. Mehta and Dharmendra G. Bhatti

Improved Google Page Rank Algorithm 535
Abhishek Dixit, Vijay Singh Rathore and Anchal Sehgal

A Pioneering Encryption Technique for Images 541
C. Jeba Nega Cheltha and Rajan Kumar Jha

A Pedestrian Collision Prevention Method Through P2V Communication 547
JinHyuck Park, ChoonSung Nam, JangYeol Lee and DongRyeol Shin

drksujatha23@gmail.com

Summarization Using Corpus Training and Machine Learning 555
Vikas Kumar, Tanupriya Choudhury, A. Sai Sabitha and Shweta Mishra

Exploring Open Source for Machine Learning Problem on Diabetic Retinopathy 565
Archana Kumari, Tanupriya Choudhury and P. Chitra Rajagopal

Diagnosis of Parkinson's Diseases Using Classification Based on Voice Recordings 575
P. Chitra Rajagopal, Tanupriya Choudhury, Archana Sharma and Praveen Kumar

Analytical Analysis of Learners' Dropout Rate with Data Mining Techniques 583
Shivanshi Goel, A. Sai Sabitha, Tanupriya Choudhury and Inderpal Singh Mehta

Discovering the Unknown Patterns of Crop Production Using Clustering Analysis 593
Dakshita Sharma, A. Sai Sabitha and Tanupriya Choudhury

Predicting the Accuracy of Machine Learning Algorithms for Software Cost Estimation 605
Chetana Pareta, N. S. Yaadav, Ajay Kumar and Arvind Kumar Sharma

Cloud Computing Research Issues, Challenges, and Future Directions 617
Dhirender Singh, R. K. Banyal and Arvind Sharma

Social Big Data Analysis—Techniques, Issues and Future Research Perspective 625
Pranjali Borgaonkar, Harish Sharma, Nirmala Sharma and Arvind Kumar Sharma

A Review Paper on Eye Disease Detection and Classification by Machine Learning Techniques 633
Neha Bharti, Geetika Gautam and Kirti Choudhary

Kernel FCM-Based ANFIS Approach to Heart Disease Prediction 643
Waheeda Rajab, Sharifa Rajab and Vinod Sharma

State-of-the-Art Artificial Intelligence Techniques in Heart Disease Diagnosis 651
Nahida Nazir, Sharifa Rajab and Vinod Sharma

Security and Privacy Issues in Big Data: A Review 659
Priyanshu Jadon and Durgesh Kumar Mishra

Stemmatizer—Stemmer-based Lemmatizer for Gujarati Text 667
Himadri Patel and Bankim Patel

drksujatha23@gmail.com

Performance Impact on Different Parameters by the Continuous Evolution of Distributed Algorithms in Wireless Sensor Networks: A Study 675
Hemlata Soni, Gaurav Gupta and V. K. Chandna

A Framework of Lean ERP Focusing MSMEs for Sales Management 683
Shilpa Vijaivargia and Hemant Kumar Garg

Multimedia Cloud for Higher Education Establishments: A Reflection 691
Anjum Zameer Bhat, Vikas Rao Naidu and Baldev Singh

Optimal Multi-document Integration Using Iterative Elimination and Cosine Similarity 699
Fr. Augustine George and M. Hanumanthappa

Secured Data Sharing in Groups Using Attribute-Based Broadcast Encryption in Hybrid Cloud 707
E. Poornima, N. Kasiviswanath and C. Shoba Bindu

An Efficient FPGA-Based Shunt Active Filter for Power Quality Enhancement 719
P. C. Naveena Shri and K. Baskaran

An Empirical Study on Potential and Risks of Twitter Data for Predicting Election Outcomes 725
Abdul Manan Koli, Muqeem Ahmed and Jatinder Manhas

An Intelligent Framework for Sentiment Analysis of Text and Emotions—Implementation of Restaurants' Case Study 733
Esha and Arvind Kumar Sharma

Author Index 743

drksujatha23@gmail.com

About the Editors

Vijay Singh Rathore is Professor at JECRC, Jaipur. He has completed his Ph.D. in computer sciences at the University of Rajasthan. He is an active member of various academic committees, including IEEE, ACM, UGC, IGNOU, RTU, UOR, IISU, and JNU. He has published over 100 papers in high impact journals and conference proceedings and has written books on *Computer Science, Information Technology and Management*. He has also filed two patents for SMILE and FOBIA.

Marcel Worring is Professor at Data Science for Business Analytics (Amsterdam Business School) and Associate Professor at the Informatics Institute (IvI). He is also Associate Director of Amsterdam Data Science (www.amsterdamdatascience.nl). He has completed his Ph.D. in shape analysis of digital curves at the University of Amsterdam. His research interests are in multimedia analytics, with a focus on the integration of multimedia analysis, multimedia mining, information visualization, and multimedia interaction into a coherent framework that yields more than its constituent components.

Durgesh Kumar Mishra received his M.Tech. in computer science from DAVV, Indore, in 1994 and Ph.D. in computer engineering in 2008. He is currently working as Professor (CSE) and Director of the Microsoft Innovation Centre at SAIT, Indore, Madhya Pradesh, India. He is also Visiting Faculty at IIT-Indore, Madhya Pradesh, India. He has 24 years of teaching and 10 years of research experience and has published over 90 papers in refereed international/national journals and conferences, including IEEE and ACM conferences. He has also served as Chairman of Computer Society of India (CSI) , CSI Indore Chapter; State Student Coordinator—Region III MP; Member-Student Research Board; and Core Member-CSI IT Excellence Award Committee. He is currently Chairman of CSI Division IV Communication at National Level (2014–2016). He had been Chief Editor of the Journal of Technology and Engineering Sciences, consultant to industry and Government of Madhya Pradesh Organizations, and Member of BIS, for the Government of India for Information Security Domain.

drksujatha23@gmail.com

Amit Joshi is a young entrepreneur and researcher with a B.Tech. in information technology and M.Tech. in computer science and engineering, and currently his research interests are in the areas of cloud computing and cryptography. He has 6 years' of academic and industry experience in prestigious organizations of Udaipur and Ahmedabad. He is an active member of ACM, CSI, AMIE, IACSIT-Singapore, IDES, ACEEE, NPA, and many other professional societies. He has presented and published more than 30 papers in national and international journals and IEEE and ACM conferences. He has also edited three books on diverse subjects and organized over 15 national and international conferences and workshops.

Shikha Maheshwari has dedicated over 10 years to educating the engineering students who will shape the country's future. She received her M.Tech. in CSE from Mody Institute of Technology and Science, Laxmangarh, in 2008. She is currently working as Assistant Professor (CSE) at JECRC, Jaipur, India. She was also recognized as a Microsoft Innovative Educator-Expert (MIE-E) in 2017. She is also a member of the editorial boards and program committees and a reviewer for numerous national and international refereed journals and conferences.

drksujatha23@gmail.com

Nuts and Bolts of ETL in Data Warehouse



Sharma Sachin, Sandip Kumar Goyal, Sharma Avinash and Kumar Kamal

Abstract Data transformation from text files to database files, relational database management systems, and distributed database management systems in recent past has emerged a vast field of data warehouse. Currently data analytics is the most appealing field for the data scientists and challenges are very big as data volume is very huge. Not only data volume is high but the speed at which data is growing annually is exponentially. Data analytics has become a tool to grow the business by forecasting, business intelligence and decision support systems. In a simplified way, data is organized in the form of database, collective databases makes the data warehouse and the technologies like business intelligence, decision support system, and data analytics make use of data warehouse for their purpose. Big data is the enhanced form of the data warehouse which consists of the cloud storage and MapReduce-based architecture which consists of the clustering of data. This survey paper will give a high-level understanding of the existing data warehouse processing mechanisms including conventional processing and the distributed processing. Existing Extraction Transformation and Loading process will be analyzed for better understanding of the sub processes of the data warehouse building process.

Keywords DWH · BI · ETL · OLAP · Optimization · DSS · Data cleansing

S. Sachin (✉) · S. K. Goyal · S. Avinash
M.M University, Mullana, Ambala, Haryana, India
e-mail: er.sachinsharma@gmail.com

S. K. Goyal
e-mail: hodce@mmumullana.org

S. Avinash
e-mail: sh_avinash@yahoo.com

K. Kamal
National Institute of Technology, Srinagar, Uttrakhand, India
e-mail: kamalkumar@nituk.ac.in

© Springer Nature Singapore Pte Ltd. 2019

V. S. Rathore et al. (eds.), *Emerging Trends in Expert Applications and Security*,
Advances in Intelligent Systems and Computing 841,
https://doi.org/10.1007/978-981-13-2285-3_1

drksujatha23@gmail.com

1 Introduction

All organizations have data and data may be in any form. It can be text, CSV, Excel, PDF, videos, and images. The whole world rely on data itself and a new domain has emerged which focused on data management. Warehouse is a place where material is stored and it is like big store which is searched when something useful is needed out of it. In the same way, data warehouse is a store where all the data is kept which includes history data and present data. There are two types of data systems.

(a) Online Transactional Processing Systems (OLTP) and
(b) Online Analytical Processing Systems (OLAP).

OLTP are real-time systems in which business transactions are performed. OLAP systems are not real time but they contain data of OLTP for analysis and retrieval. Data warehouse and OLAP are correlated words and are used interchangeably. A organization have many domains and verticals which can be Customer Relationship Management (CRM), Enterprise Resource Planning, plain text/CSV data from various departments, Online Transactional Processing System data and data from various other sources. The data formats are different and data sources vary from one platform to another. Data warehouse uniform the data available from various data sources so that data analysis can be done in an effective manner. Data warehouse and ETL mechanism is shown in Fig. 1.

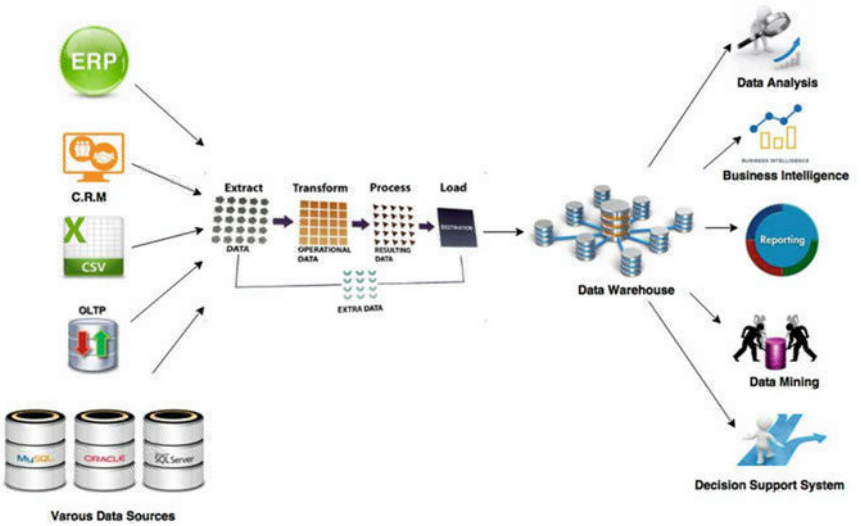


Fig. 1 Data warehouse and ETL

drksujatha23@gmail.com

2 Related Work

Authors in [1] have put light on the importance of query optimization for cloud data warehouse. The outsourcing methodology is used for cloud computing so as to customize the processing and infrastructure power as per the requirement. Due to cost effectiveness and flexibility of the cloud, many organizations have been migrated their data applications to the cloud framework. The major challenge is data processing time which should be improved and optimized in the cloud for data warehouse. The legacy cloud optimization techniques are still required for cloud data warehouse which includes indexing, partitions, and views. The cloud systems are having multiple nodes which are distributed on the cloud network based cluster wise. The query requesting heavy data volume from cloud data warehouse will check into different nodes and needs frequent inter-node communication. The performance is directly linked with inter-node communication. Higher is the inter-node communication to extract the data, more performance issues arise. Authors in [1] have suggested improving DWH performance in cloud. Authors in [2] point out ETL optimization by improving its execution. The different phases of ETL and their execution order have been reviewed. ETL tool functions have been identified by the authors as source data identification for relevant information, extraction of the source data, information dissemination, converting data coming into desired data format which may be referred as transformation, data cleansing based upon rules followed by data loading into fact tables which will lead to the data marts and data warehouse. Authors in [3] have put a light on the better understanding of the data warehouse and the area of application. The authors have detailed the concept of data warehouse, relational databases, its scope, enterprise systems, decision-making, and Hadoop platform as an approach for data integration. Apart from transactional data, there is a lot other data which can be used for fetching useful information. The unstructured data is hard to store and analyze. The data available on social networking sites, data generated by educational institutions, research data of science and technology and other semi-structured data may be analyzed for decision-making. Authors have tried to make use of Hadoop and big data in decision support system for Education which has emerged a new area for researchers.

Authors in [4] have enlightened the crucial role of ETL tools in the data warehouse building process. Source to target data mapping has been achieved through SQL queries. Authors have given an SETL tool which is an Extraction, Transformation and Loading tool based on SQL queries. SETL tool proposed in [4] interacts with the repository of the DBMS and make transformation operations on the source data so that it can be mapped with target data. The temporary database objects have been created and transformation operations have been performed on it. Upon getting fact data dimensions from the staging area, the temporary objects have been purged. Data quality is a major concern in the data warehouse building process which generated many errors and sometimes results into halt in the DWH System. Authors in [5] have tried to address the data cleansing issues in the data warehouse.

drksujatha23@gmail.com

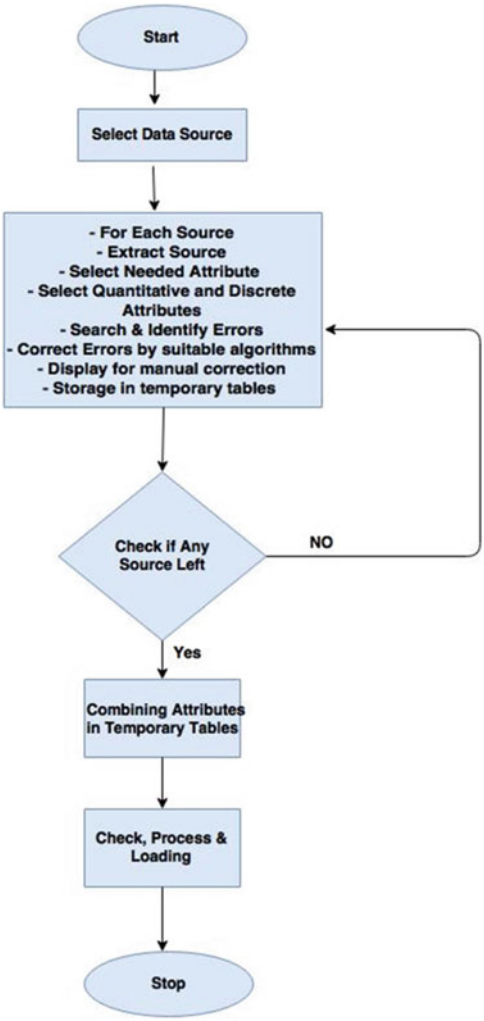


Fig. 2 Enhanced data cleansing algorithm [5]

Authors have provided an algorithmic-based approach for error detection and correction in the data warehouse process. The ultimate goal is to provide error free data warehouse in which all the data quality errors like data duplication errors, lexical data issues, data format issues, etc. The main data quality characteristics are completeness, validity, precision, accuracy, data integrity, accessibility, conformance, and timelines. The algorithm as provided by authors in [5] is shown in Fig. 2.

Authors in [6] have put in their effort for workflow optimization to minimize the cost for real-time data warehouse systems. Authors have proposed algorithm for workflow transformation which is improving the communication cost and computational cost during data warehouse processing. Real-time data warehouse needs

drksujatha23@gmail.com

stream workflows where stream workflow consists of four tuple system (A, B, C, D) where A is a set of finite n vertices (a_1, a_2,…, a_n), B is set of finite m vertices (b_1, b_2,…, b_m), C is a set of k directed edges and D is a set of one directed edge. The DWH cost execution depends upon various factors like virtual machines performance, data centers connectivity and other factors. Authors have provided an algorithm for cost effectiveness which optimizes the workflows of data warehouse. Authors in [7] have enlightened the role of ETL process modeling and organizing. Conceptual design for ETL is the first stage in the ETL process and authors have proposed real-world ETL process modeling and organization based upon the their experience. A new framework is proposed which takes four inputs and graphical conceptual ETL model is the output of this process. The rejection in the ETL process is being handled and ETL modularization is been exposed of. Necessary inputs have been provided to obtain the ETL conceptual model which is represented graphically. The organizing of ETL process is designed in such a way that ETL process is divided into six sub-processes. There is a reject module and the normal modules. The technical check, semantic check, standardization and cleaning, mapping and processing of the data is performed in a sequence and the reject module is associated with all the modules such that the bad data can be collected, analyzed, and processed in an effective manger.

Call data record [8] is the record of every telecom subscriber who made a call. Its having various data fields which contains information regarding the customer including location detail, IMEI (International Mobile Equipment Identification), IMSI (International Mobile Subscriber Identity), cell information (Tower from which made a call), date and time of making a call and various other information. Telecom sector is growing at a rapid rate and the information of the subscribers is enormous. CDR (Call Data Record) is the key component by which we can analyze the many parameters about customer. Authors in [8] have suggested methodology to process call data record and analyze the market share of the telecom equipment manufacturers. Authors have given a light on processing huge data volume of the CDR to analyze the IMEI based data so that mobile handset company share can be depicted and based upon data, further decisions can be made. Authors in [9] have provided a way ahead for researchers on the point that huge data processing can lead to the compromise on the security aspect of the data and data can be compromised. Adding security layer to the data warehouse is a big challenge as it will result in the performance loss during data processing. Data volume is very large and as a result, security can be on stake. Authors have proposed security driven architecture for big data warehouse. A warehouse server concept has been introduced which act as the enforcement for data security rules while sharing data with the end users. The architecture of security driven data warehouse is in Fig. 3.

Authors in [10] have observed that ERP implementation of majority of projects has been delayed due to various reasons. Authors in [10] has detailed database and business intelligence tools belongs to applications, products, and systems. The ultimate goal is to reduce the time taken by SAP tools to extract, transform, and

drksujatha23@gmail.com

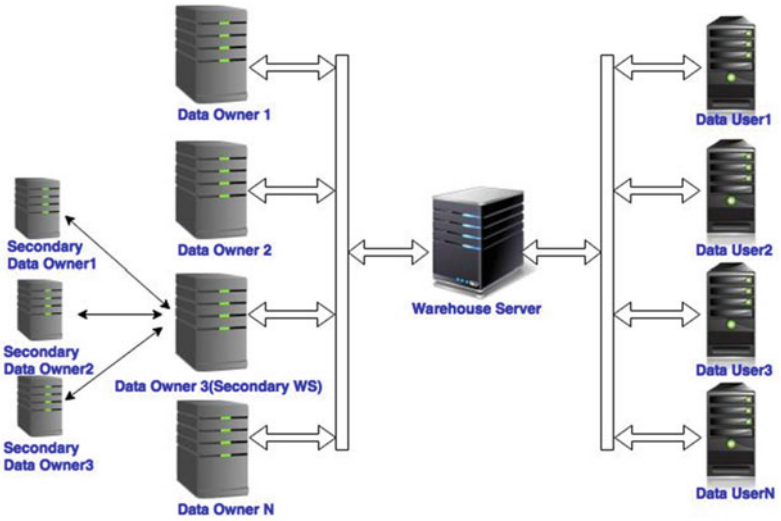


Fig. 3 Security driven big data warehouse architecture [9]

loading of data. The success of ERP depends upon the effective ETL part and efficient data model. SAP BI is the module provided by SAP which takes care of the ETL and reporting part. Performance has been kept at top priority and focus has been given on the ETL part and reporting run time. The source data can be SAP based and Non-SAP based. In case of SAP data source, SAP interface can be used and similarly in case of non- SAP data source, separate interface is used for extraction. Authors in [10] have come to conclusion that the performance time is reduced in the ETL and reporting in case of SAP tools for ERP implementation. Authors in [11] have described ETL as a process of fetching data from source, change it to the format required in final tables and then loading the same into the destination tables. By making use of different technologies—Oracle, Java and XML; authors have proposed a new model called Hyper-ETL. The aim is to eliminate the metadata mismanagement. To optimize sale promotion services, SQL code generation is used. Storage space optimization has been achieved by split the columns after merging of different tables. Hyper-ETL design steps are as Source data extraction, User requirement based table creation with required attributes, data transformation for operations and generate meta-data for XML file. Authors have presented experimental results which show that a significant improvement has been made by using Hyper-ETL. Two million records have been processed in the relevant tables in one hour and fifty seven minutes. As detailed in [12], authors have detailed the new concept of extraction, transformation, loading and retrieval optimization technique.

drksujatha23@gmail.com

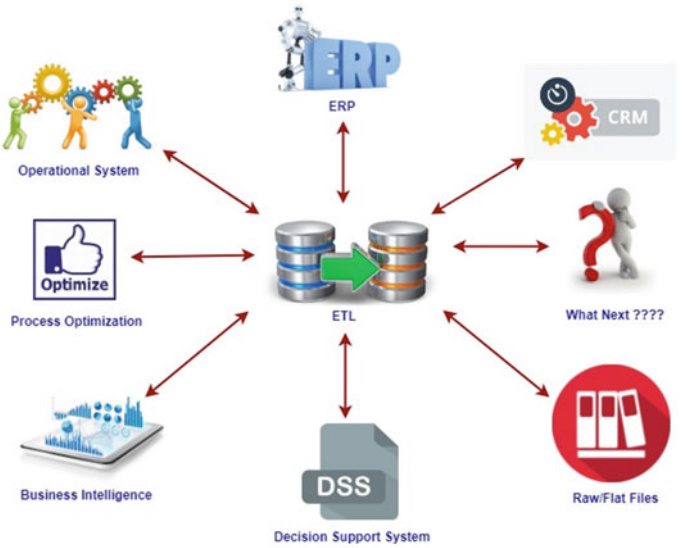


Fig. 4 Nuts and Bolts of ETL

3 Nuts and Bolts of ETL

Nuts and bolts of ETL comprise of many components including operational system, enterprise resource planning, CRM, processing raw files, decision support system, business intelligence, process optimizations, etc. ETL is in center of many systems and remains a key component for enterprise data warehouse solution. Despite the fact that big data technologies have evolved in recent years, importance of ETL and its role in constructing a data warehouse is inevitable and is specified in Fig. 4.

4 Conclusions and Future Work

Data has become the major challenge for every industry and data processing has been evolved in many forms. The volume of data is so high that it is termed as big data, and big data consists of the cloud network and standalone infrastructure. Data warehouse is not merely a set of data marts but it is acting as an interface to the various domains at the same time and is in use for various purposes including DSS, Business intelligence, forecasting, obligatory requirements and many more. In this paper, we have thoroughly studied the ideas given by various researchers in the area of data warehouse optimization when it comes to the large data sets of high volume data. Following is the list of ideas given by various researchers in the area of Big data processing:

drksujatha23@gmail.com

- Cloud DWH query optimization
- ETL Optimization Process
- Decision Support System in Education Using Big Data and Hadoop
- New Tool for ETL Processing
- Improved Data Cleansing technique in Data Warehouse
- Real-Time Big Data Process Optimization using Workflow Transformation
- ETL Process modeling and organizing
- Call Data Record Analysis to Identify the Market Shares
- Security Driven Big Data warehouse Architecture
- ETL and Reporting Optimization

Optimization of the data warehouse by researchers has been studied which covers different parameters involved in it like security, Cloud, query optimization, real-time processing, report optimization and many more. There is a need of overall optimization model which should not be processed individually but should be in built and act as natural way of DWH processing. All the aspects and every kind of industry should be covered in a uniform optimization model. Also, we feel that as technology is emerging day by day, therefore future technology mix should be consider which also maintains analytics flexibility provided by the traditional RDBMS. There is lot of scope of work to be done in the area of DWH optimization with uniform model which covers map reduce, no sql, hadoop, spark systems whereas maintaining the querying capabilities like RDBMS has been offered.

References

1. Abdelaziz E (2015) Optimisation of the queries execution plan in cloud data warehouses, pp 129–133
2. Anand N, Kumar M (2013) anand2013. In: Modeling and optimization of extraction-transformation-loading (ETL) processes in data warehouse: an overview
3. Bondarev A, Zakirov D (2016) Data warehouse on Hadoop platform for decision support systems in education
4. Chen Z, Zhao T (2012) A new tool for ETL process. In: Proceedings of the 2012 international conference on image analysis signal processing IASP 2012, pp 269–273
5. Hamad MM, Jihad AA (2011) An enhanced technique to clean data in the data warehouse. In: Proceedings of the 4th international conference on developments in eSystems engineering DeSE 2011, pp 306–311
6. Ishizuka Y, Chen W, Paik I (2016) Workflow transformation for real-time big data processing
7. Kabiri A, Chiadmi D (2012) A method for modelling and organazing ETL processes. In: 2nd international conference on innovative computing technology INTECH 2012, pp 138–143
8. Maji G, Sen S (2015) A data warehouse based analysis on CDR to depict market share of different mobile brands. In: 2015 annual IEEE India conference, pp 1–6
9. Mukkamala R (2016) Privacy-aware big data warehouse architecture
10. Parul, Nawab S, Teggihalli S (2015) Performance optimization for extraction, transformation, loading and reporting of data. In: Global conference on communication technologies GCCT 2015, no. Gcct, pp 516–519

drksujatha23@gmail.com

11. Prema A, Pethalakshmi A (2013) Novel approach in ETL. In: Proceedings of the 2013 international conference on pattern recognition, informatics and mobile engineering PRIME 2013, pp 429–434
12. Sharma S, Kumar K (2016) ETLR—Effective DWH design paradigm. In: Proceedings of the international conference on data engineering and communication technology. Springer, pp 149–157

drksujatha23@gmail.com

User Identification Over Digital Social Network Using Fingerprint Authentication



Devender Dhaked, Surendra Yadav, Manish Mathuria and Saroj Agrawal

Abstract Today, the entire world of web communication is governed by Internet. Through Internet data is transformed digitally. The main benefit of transferring the information digitally is that an authenticity is managed between both the sides (Sender and Receiver) for making a reliable transformation over the Internet. The main role of security is concerned through this type of web communication. Now social networking sites are playing a vital role in our life for sharing our live events through different Medias such as audios, Images, and Video files, etc. But the problem which is arising during the communication on social sites is: a pretender can easily access other's account information like picture or any other detail because on the social networking sites it is easy to copy. Nowadays, the question arises: "how to verify the user's real unique digital identity on the social network?" In this research, the measure concern is about the security over the social networking sites. The proposed research work is the solution that can secure the privacy of a Digital Identity with the use of Digital Watermarking Technique. This method works on the concept of Digital Fingerprint. Where for watermark image Digital Fingerprint is used which is embedded in the original image using Watermarking technique discrete wavelet transform (DWT).

Keywords Social identity · Social networking sites (SNS) · Digital communication · Digital fingerprint watermarking · Color image digital watermarking · Discrete wavelet transforms (DWT), etc.

D. Dhaked (✉) · S. Yadav · M. Mathuria
Department of Computer Science, MACERC, Jaipur, India
e-mail: devenderdhaked003@gmail.com

S. Yadav
e-mail: syadav66@gmail.com

M. Mathuria
e-mail: manish.4598@gmail.com

S. Agrawal
Department of Computer Science, JECRC, Jaipur, India
e-mail: sarojagr708@gmail.com

© Springer Nature Singapore Pte Ltd. 2019

V. S. Rathore et al. (eds.), *Emerging Trends in Expert Applications and Security*,
Advances in Intelligent Systems and Computing 841,
https://doi.org/10.1007/978-981-13-2285-3_2

drksujatha23@gmail.com

1 Introduction

Now in the present world Communication is being held through web digitally which is a generation of Digital Information Technology. Since 2000, human dependency on computer for socializing purpose is growing rapidly. The improvement of innovative technologies focuses resting on the accessibility of digital in sequence toward the spectators. The idea of release resource is suitable through the inventors of information knowledge; and so, the requirement for digital information has furthermore increased. The concept of release resource not only provides liberty for persons to modify the data according in the direction of their own needs, although it in addition opens the track for the inappropriate use of freedom.

If a person receives an exclusive ID by the unique ID certificate (UID) as a citizen of a country, then the question comes: "Why it cannot be issued for being the user of any social networking system?" But, it will not be acceptable for the reason that, social networking is seen as components of living where user is capable of liberally contribute to their thoughts. The main purpose of this research investigate document is to offer universal understanding regarding the significance of digital uniqueness on public networking systems [1].

2 Uniqueness of User

2.1 Digital Identity

For identifying an owner digitally a Digital identity is provided. Mainly, data on mobile devices or computers can be shared digitally through the network, so digital information is required for finding the owners of digital content [2].

2.2 Social Identity

It is the identity of human beings on the social networks. Social networking manages communication either digitally or manually. Users require social identities to present some of their followers known as; people belonging to their communities need an identity (or designation). The connectivity of users to their social group is called social networking. Members of social networking can socially work in partnership on an event in their society [3].

drksujatha23@gmail.com

2.3 Fulfilled Points of Discussion

Digital identification uses only to represent a unit's digital presence activity, it is not related to personal user personal information. Social Identity is associated not only with the individual's information but also the information about user's public relationships. Various social identities communicate to enduring social categories, like as ethnicities, religions, or nationalities. In existing scenario both, the user gets recognition to attend anything, but there is no organization to verify the uniqueness of the any human. Both, i.e., Digital Identity and Social Identity are replaced by new keyword which is "Digital Social Identity" key feature of this research, as the objective is to present major problems of Social Networking Users with their privacy.

3 Digital Social Networking

Working Concept of any social networking is to offer a public stand for everyone to deal out their thoughts throughout the use of social media. It adheres to the open impression, where the users can use their appliance and contribute to their data lacking cost and donation. Its general directions given through several public networking website are:

3.1 Groupware

Groupware is the prime concept of digital social networking. This is an essential thing of several social networks that compose network within organization. Essentially, a social network, itself is a compilation of several personal organization. These systems are generated through a web application and a web tools. In further terminology, using groupware, groups are developed well maintained. Groupware are categorizing in two categories: Unprotected Groups and Protected Groups. Unprotected Groups are oldest process of constructing group and then inviting your friends to join the group. This kind of group facilitates substance sharing, with some limitations, where, all the public information can be seen to all the associated users. Other side in protected Groups, advanced groupware facilitates creating individual groups and restricting accessibility for human being. This kind of groupware offers progress distribution facilities with describing content as ridiculous, but still requires various laws of those policies that are necessary to offer security measures and assurances.

drksujatha23@gmail.com

3.2 Active Wall (or Sharing Wall)

These SNS maintain a proposal for the respective user in order that users can convey our opinion by means of their supporters (or joined group). Move ahead digital societal networking site helps the customer detect their digital substance such as World Wide Web documents, picture, audio, and video files, etc. An Active Wall Group member is a means of communication where the sharing Wall essentially reflects the in particular action of several users on the social network.

4 Various Security Issues for Identity

The security of any deal on the Internet is mandatory; the main concept behind security is authentication with the use of "username "and "password". This is the basic tool for authentication, but for some time, special digital system is required on a special network of authentication. Before discussing security issues about public network, one should understand "What is required for digital social identity?" [3].

4.1 What Is the Importance of Digital Social Identity?

The same as mentioned on top of, societal identification is defined by social identity on social networking websites. It is a combination of digital as well as public uniqueness. Once a creator has created a login account on the digital social networking website, the website requires few private information about user name, user email, phone number, location (address), etc., as it is a fixed step to creating an account. It is fine when someone shares their personal information with the social networking site to create an account, but how to confirm about one's security and that, this information will not be shared with others. And if someone can see the personal information of a person, then it nullifies the idea of unique identity, as anyone can create an account using the personal information of others on the Internet. Given the individual specificity of any user on the social network, this is a big problem. This is the right time, when the IT industry should meeting point on digital social uniqueness. A solution to the problem is a personal factor that can provide personal information to the individual with specific information toward the abuser and it must not be worth sharing. Research implementation gives the digital fingerprint result for digital unique identification with personal information [3–5].

drksujatha23@gmail.com

4.2 Privacy Issues

Any person can easily create duplicity account. User's Profile and Personal Information can be easily hacked. Fake users show off like the real abuser with using personal data and image. Nobody be capable to blame any person. There is no pleas registration system, condition somebody wishes to building block your counterfeit accounts. If a personal certificate and picture is derivative through fraud person also they make use of it on private documents and as profile images; after that there is section to prove to identity of the original user [6].

5 Security Methods

Security is a progression for authentication that is used to verify identity. Digitalized certification contains recognition procedure that can detect the source of information. In case of digital data communiqué, there are two accepted techniques in digitalize authentication.

5.1 Digital Signature

It is a statistical procedure that validates the dependability and integrity of a message, software or digital document. This is basically a digital signature which secretly authenticates the information sender. Digital signatures are working on public key cryptography. Encryption is used on the sender to mark the significance through the personal explanation. And by the side, i.e., the recipient is capable to confirm the autograph with the sender's open key.

5.2 Digital Watermarking

Watermark is a very old technique used for copyright infringement and bank note authentication. But now, any digital watermarking technique can be used to mark on any digital signals like as audio frame, pictures, videos, and text or "Digital watermarking approach achieved universally used to point any information within digital media for the authentication of ownership" Or "Digital watermarking approach used for copyright protection to hidden the user data within digital multimedia" Watermark is a digital identity technique that holds the same information as the user. Digital watermarks can be used to verify the authenticity or integrity of the media. It is believed that the theory of digital watermarking allows information to be influenced but does not allow the abusive quality of the data [7].

drksujatha23@gmail.com

5.3 *Fingerprint as Watermark*

The essential idea of this follow a line of investigation is to emphasize upon the uses fingerprint picture as watermark image. Fingerprint be use in the direction of recognize a human being specially. Therefore, a digital fingerprint image can be used as a watermark. The source of fingerprint images for practical analysis is author's own scan impressions of finger actually. The same as we discussed on top of, digital fingerprinting is an imperative process that is capable of simply be use to recognize individual identities, although it must exist hidden. Unseen or hidden watermarking is obtained by discrete wavelet transform [8]. The functionality of DWT techniques has more robust and secure digital watermarking. Image Watermarking of gray scale image is researched by many organizations but watermarking of colored image is still required to be deep research for good quality image. DWT has its advantage of robust watermarking. DWT techniques are also very useful for colored image because in the present scenario, color image is really simple headed for imprison, reproduction, edit, and allocation. Therefore, this investigates generally concentrate on top of most excellent copyright protection and authentication using new feature of Digital Watermarking [9].

Figure 1 shows a building block diagram of fundamental move toward to using fingerprint in the form of watermarks is to display the algorithm and their steps embedded watermark in a colorful image using the DWT (discrete wavelet transforms) middle block called "watermark editing" Where IDWT (Inverted Wavelength

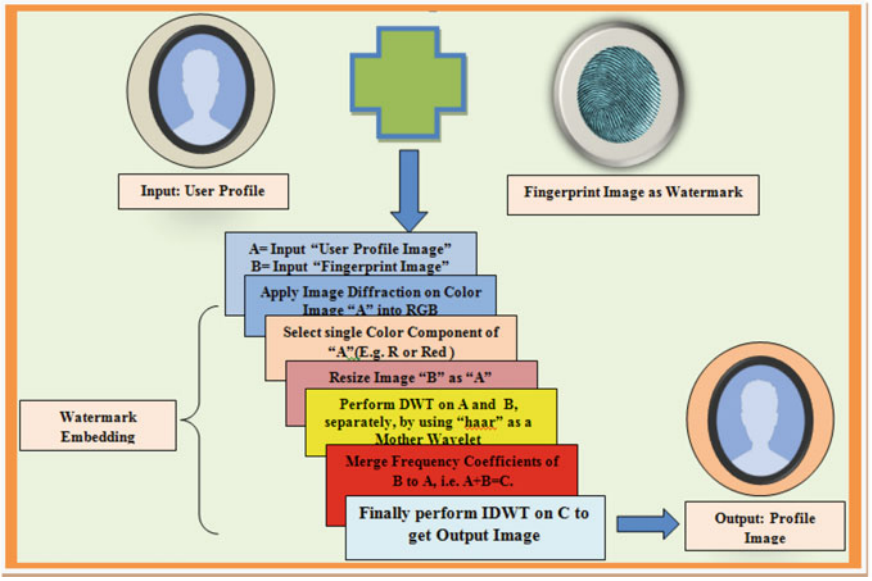


Fig. 1 Block diagram of color image digital watermarking selecting fingerprint image as watermark

drksujatha23@gmail.com

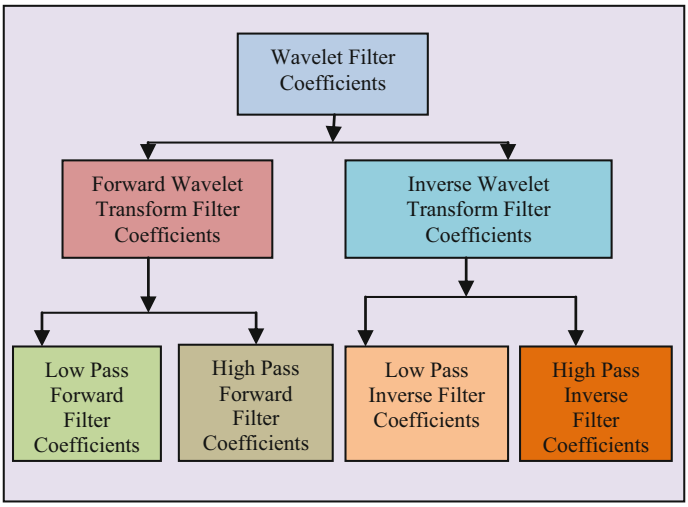


Fig. 2 Wavelet composition of high and low filters [10]

Transform) is used to reproduce the image of the profile from frequency coefficient Entrance is used [10]. Two-dimensional Discrete Fourier or Cosine Transform have been represented in the advance Discrete Transform function like F[u, v], which is a function of fully spatial frequency u or v. No straight information regarding pixel or spatial variables is been given likewise, Discrete Wavelet Transformation use additional types of necessary functions, i.e., Daubchies, Haar. This original work is furthermore identified like Mother Wavelet. In this investigation, Haar is use like a Mother Wavelet [11, 12] (Fig. 2).

6 Experiment and Result

This research also includes the implementation of proposed schema on MATLAB. The number of images was used as a practical image and finally measured the corresponding affection of the image.

6.1 Select Profile Input Image

To test the efficiency of algorithm, this experiment has been conducted on a color image "Rangoli.jpg" are used (Fig. 3).

drksujatha23@gmail.com



Fig. 3 Original input image for watermarking: Rangoli.jpg

6.2 Select Fingerprint Watermark Image

See Fig. 4.

6.3 Watermarked Profile Picture Image

See Fig. 5.



Fig. 4 Watermark: Fingerprint.jpg (scan impression of author's own finger)



Fig. 5 Watermarked image: W_Rangoli.jpg

drksujatha23@gmail.com



Fig. 6 Recovered watermark: RW_Fingerprint.jpg (processed fingerprint impression of author's own finger)

6.4 Recovered Watermark

See Fig. 6.

6.5 Image Quality and Similarty Measurement

In the Digital Watermarking Peak Signal-to-Noise Ratio (PSNR) Value decides the Image quality. When PSNR value comes under the range 30–50, Good quality of image is creating.

PSNR: It is Peak Signal-to-Noise Ratio that reflects the ratio between the maximum possible power of a signal and power of corrupting noise. It also affects the originality of the signal. The PSNR is usually expressed in terms of the logarithmic decibel scale. It is commonly used to check quality construction of loss compression. Here, MAXI represents the maximum possible value of the pixel in the image.

The PSNR calculation (in dB) is defined as

$$\begin{aligned}\mathbf{PSNR} &= 10.\log_{10}\left(\frac{MAX_I^2}{MSE}\right)\\ &= 20.\log_{10}\left(\frac{MAX_I}{MSE}\right)\\ &= 20.\log_{10}\overline{(\mathrm{MAX_I})} - 10.\overline{\log_{10}}(\mathrm{MSE})\end{aligned}$$

Coefficient of Correlation (CoC): Correlation coefficient is also used for calculating the image quality. Where, x_i and y_i are strength values of ith pixel in first and second image respectively. Also, x_m and y_m are mean intensity values of first and second image respectively. Correlation coefficient finds out the similarity between

drksujatha23@gmail.com

Table 1 Image quality testing

Image type	Peak signal noise ratio and correlation coefficient		
	Compared images	PSNR	Correlation coefficient
Profile picture	Rangoli.jpg and W_Rangoli.jpg	34	0.9827
Watermark image	Fingerprint.jpg and RW_Fingerprint.jpg	13.2346	0.8101

original image and watermarked image, which would be within 0–1 for determining the similarity between the images.

$$r = \frac{\sum_i (x_i - x_m)(y_i - y_m)}{\sqrt{\sum_i (x_i - x_m)^2} \sqrt{\sum_i (y_i - y_m)^2}}$$

When correlation coefficient value $r = 1$ then both the images are completely similarly, when $r = 0$ then both the images are completely uncorrelated and when $r = -1$ are completely anti-correlated [13] (Table 1).

7 Future Scopes

This research work is very interesting and efficient. People are more comfortable using biometric security, there are public concerns over its validity. It refers to the metrics related to the human characteristics, particularly the physical and behavioral aspects. Every individual is unique and carries a separate identity in the form of traits like fingerprints, hand geometry, iris recognition, voice, etc. It can be used in many cases like Biometrics Technology, Face Recognition Technology, Biometric Time and Attendance System in Offices, Fingerprint Recognition in Smartphone's, Wearable Gadget monitoring the brainwaves of the body. The core advantages of unique fingerprint identification are **easiness of scanning**. Technically, this work can be further improved by using the other algorithms in which used parameters can give the less PSNR values for best image quality.

8 Conclusions

Each and every person around the world prefers having their own unique identity for use on any social network. Unique presence provides them the confidence to distribute their individuality information independently without accommodation on some protection matter. In this research paper fingerprint watermarking is performed by using DWT approach. The reviewed research about the fingerprint

drksujatha23@gmail.com

watermarking is very much appreciable thus many research articles found on this topic. The watermarking is a only technique to avoid the misuse of image documents. DWT has advantage of robust watermarking. The aim of this research is to emphasize that unique identity should be provided to individuals for their use on the digital social networks. This research is focused on color image digital watermarking. In this research process conclusion that are come out from the experiment is that the watermarked image is identical and after recovery we can easily identify the watermarking.

The working process of this following research of investigation concludes that if the social networking websites start providing an individual's identity for use on any digital social networks, it can easily take access on the following: It easily controls the privacy settings of individual user's. It protected the virtual identities of user on different social network site. It can finally manage the unfaithfulness of any data. It can attain complete confidence of individual's user. Its Unique Identification may provide single integrated environment. Facilitate user to find out their elderly friends. Fraud can easily find out. Fake users will not be able to generate several duplicity accounts, for the reason that it will require exceptional fingerprint picture for identification and verification.

References

1. Warburton S, Hatzipanagos S (2012) Digital identity and social media. IGI Global. ISBN: 9781466619159, Released July 2012
2. Camp LJ (2004) Digital identity. IEEE Technol Soc Mag 23(3)
3. Hornsey MJ (2008) Social identity theory and self-categorization theory: a historical review. Social Pers Psychol Compass 204–222. Blackwell Publishing Ltd.
4. Van Lange PAM, Kruglanski AW, ToryHiggins E (2012) Handbook of theories of social psychology, vol 2. British Library. ISBN: 978-0-85702-961-4
5. Sullivan C (2011) Digital identity an emergent legal concept. University of Adelaide Press
6. Fitzgerald BF, Xiaoxiang Shi S, O'Brien DS, Gao F (eds) (2008) Copyright law, digital content and the Internet in the Asia-Pacific. Sydney University Press, Sydney, N.S.W. ISBN: 9781920898724
7. Jiang X (2010) Digital watermarking and its application in image copyright protection. In: International conference on intelligent computation technology and automation. IEEE
8. Srinivasulu Reddy D, Varadarajan S, Giri Prasad MN (2013) 2D-DTDWT base image denoising using hard and soft thresholding, vol 3, no 1, pp 1462–1465
9. Wei Y, Hao Y, Li Y (2009) Multipurpose digital watermarking algorithm of color image. In: Proceedings of the 2009 IEEE international conference on mechatronics and automation, August 9–12, Changchun, China
10. Bhargava N, Mathuria M (2012) Color image digital watermarking. In: Springer proceeding of international conference ICERECT series: lecture notess in electrical engineering, vol 248, in press
11. Bhargava N, Sharma MM, Garhwal AS, Mathuria M (2012) Digital image authentication system based on digital watermarking. In: IEEE conference publications of ICRCC, pp 185–189

drksujatha23@gmail.com

12. Nuruzzaman M (2005) Digital image fundamentals in MATlAB. AuthorHouse 08/23/05. ISBN: 1-4208-6965-5 (sc)
13. Jen EK, Johnston RG, The ineffectiveness of correlation coefficient for image comparisons. Research paper prepared by Vulnerability Assessment Team, Los Alamos National Laboratory, New Mexico

drksujatha23@gmail.com

A Review on Machine Translation Systems in India



Shikha Maheshwari, Prashant S. Saxena and Vijay Singh Rathore

Abstract Translation is the undeniable necessity to abrogate the correspondence hindrance. The obstruction may occur while knowing distinctive languages and prevent from sharing the information. This research paper provides a detailed review of Machine Translation (MT) systems developed for the Indian language set. It additionally gives a thought in regards to the approaches and evaluation techniques used for translation. From this paper, a researcher can have a look in regards to the work done upgrading the work from where it stops. A few systems are created for general domain, whereas others are for a particular domain like authoritative document translation, news translation, youngsters' stories, climate portrayal and conference papers, etc., and still some languages require more considerations.

Keywords Machine translation (MT) · Indian languages · English-Indian languages

1 Introduction

Although started more than a decade ago, machine translation is an emerging research is in Natural Language Processing (NLP) especially for Indian language sets. Currently, a good number of government along with private sectors are working for the developments or enhancements of the full-fledged MT system specially for a set of Indian languages as there is growing demand for high-quality automatic translation. Many attempts have been made for improving as well as

S. Maheshwari (✉) · P. S. Saxena
Jaipur National University, Jaipur, India
e-mail: shikhamaheshwari6583@gmail.com

V. S. Rathore
Jaipur Engineering College & Research Centre, Jaipur, Rajasthan, India
e-mail: vijaydiamond@gmail.com

© Springer Nature Singapore Pte Ltd. 2019

V. S. Rathore et al. (eds.), *Emerging Trends in Expert Applications and Security*,
Advances in Intelligent Systems and Computing 841,
https://doi.org/10.1007/978-981-13-2285-3_3

drksujatha23@gmail.com

developing an MT system for English to many Indian Languages along with its reversal process. Literature shows that the majority of work in MT was done in Hindi and Dravidian languages.

2 Literature Survey

A brief description of the major machine translation developments in Indian context is presented in this manuscript.

2.1 Anglabharti

R. M. K. Sinha et al. [1, 2] proposed ANGLABHARTI, a multilingual MT for English to specific Indian languages, especially Hindi, which relies on a pattern governed methodology. The methodology in this MT framework is superior to transfer approach and lies underneath the Interlingua approach. At the primary stage, a pattern directed parsing is generally performed using the English as a source language, which produces a pseudo-target relevant to a wide set of Indian languages. Word sense ambiguity sentence likewise is determined by various semantic tags. With an end goal to change the pseudo-target language into the relating target language, the framework utilizes a different text generator module. After performing correction over all ill-formed target sentences, a postediting bundle is used. Even though it is broadly useful framework, it has been connected only for the work in the public health domain at present. The system is implemented for English-Hindi language called AnglaHindi, a web-enabled (http://anglahindi.iitk.ac.in) for obtaining good domain-specific results for various health Campaigns, and also helped in the successful translation of many pamphlets and medical booklets. Currently, further research work is going on to implement this framework for English-Telugu/Tamil translation [2, 3].

2.2 Anglabharti-II

To solve the disadvantages of the previous system, ANGLABHARTI-II MT architecture was introduced. Here, a Generalized Example Based (GEB) and Raw Example Based (REB) were used as hybrid to improve the translation performance. This system invokes the rule-based after attempting a perfect match in GEB and REB. Apart from this, machine governed pre-editing and paraphrasing steps are included in this proposed system. For achieving more robustness and accuracy, the architecture of this system supports pipelining of sub-modules [2, 3].

drksujatha23@gmail.com

Currently, this technology is bifurcated under eight different sectors across the nation with the intention to develop MT for English to 12 Indian regional languages.

2.3 Anubharti

It is an EBMT-based hybridized approach along with some primary grammatical analysis. In Anubharti, the traditional approach has been improved to reduce the necessity of a large database of examples, which is carried out by generalizing the sentence elements followed by changing it with abstracted form and identifying its syntactic groups, from the given raw examples. Matching of the input sentence with abstracted examples is then performed [3].

2.4 Anubharti-II

This system was designed for translation of Hindi to any other Indian languages, to overcome the drawbacks of the early architecture for different paradigms with a varying degree of hybridization, using a generalized hierarchical example-based methodology. This system was successfully implemented with good results by IIT, Kanpur [1, 3].

2.5 Anusaaraka

AksharBharati et al. proposed Anusaaraka, with the only defined goal of translation from given Indian language to another. Although the translated output is not grammatically correct always but still reader can understand. Therefore, it enables a user visiting a site which is in a language he does not understand, but still, he can run Anusaaraka and read and understand the topic context. It successfully works in Telugu, Kannada, Bengali, and Marathi to Hindi [1]. At present is system is successfully implemented for translation of children's stories and five regional languages into Hindi. This system consists of two modules, namely core Anusaaraka, based on language knowledge and domain-specific module, based on statistical as well as world knowledge. The main idea to bifurcate the translation load such that the machine carries out the language based analysis of the text followed by the knowledge-based analysis to be performed by the reader itself.

Presently, LTRC at IIT Hyderabad is implementing this system architecture with the main focus on English-Hindi translation. While preserving the basic principles of original Anusaaraka, it uses the modified version of super tagger based on XTAG for performing analysis on the source text along with the light dependency analyzer.

drksujatha23@gmail.com

The merit of this system is on completion of source text analysis; the user may read the generated translation and can always look for simpler output in case of failures or wrong output.

2.6 AnglaHindi

Developed at IIIT-Kanpur, this system is a pseudo-interlingua rule based for translating English to Hindi. Besides using all functionality of AnglaBharti, abstracted example based is also used for doing the translation repeatedly appearing noun along with verb phrases. Up to a length of 20 words, this system has given 90% acceptable translation results for a unique set of simple, compound and complex sentences [4].

2.7 MaTra

Funds support by (TDIL), a human-assisted translation system for English as source languages to Indian languages as target languages, with the main focus on Hindi and based on transfer approach while using frame-like structures [5]. The objective is to empower users to visually examine the system providing disambiguation of information, thus permitting the output as only correct translation. It is completely automatic for producing intermediate translations for its end users, and can be categorized as translators, content provider, and editors. The lexical approach of MaTra is for general purpose use, but its application can be effectively identifiable in the domains of annual reports, news, and phrases of technical use.

2.8 Mantra

It is developed by CDAC, Bangalore in 1999 under the supervision of Dr. Hemant Darbari [1] with the intention to carry out translations for the domains of gazette notifications related to the appointments by the government and parliamentary proceeding reports between English-Indian languages and vice versa. In this system, source and target language grammars are represented using LTAG formalism while retaining the structure of input word documents throughout the translation. Using Mantra MT, language pairs like Hindi-English and Hindi-Bengali translation already have started for RajyaSabha [3, 6].

drksujatha23@gmail.com

2.9 Shiva and Shakti

Both of these projects are developed concurrently by IISc, Bangalore and IIIT, Hyderabad and are designed for generating translation text without taking much time. SHAKTI MT system combines statistical approach on rule-based approach in which the set of rules on the basis of linguistic nature followed by statistical approach makes an attempt to infer linguistic information. Currently, Shakti is in use for three language set as output, viz., Hindi, Marathi, and Telugu [7]. SHIVA MT is based on example-based approach. Some modifications of this system also use semantic data for experiments, trials of getting the user feedbacks, and therefore publically available [8].

2.10 Anuvadak English—Hindi MT

It is a general purpose tool developed by Super Infosoft Pvt. Ltd., Delhi and contains inbuilt dictionaries for limited domains with postediting support features. The system has the facility to translate the source word present in the target language if it is not found in the lexicon. The system is available for public use easily usable on Windows Operating Systems [3, 9].

2.11 English-Hindi SMT System

Developed by IBM India Research Lab [3], this system is based on statistical machine translation approach between English and other Indian language set. IBM research has also developed several NLP related tools for the quality improvement of translations using SMT systems. Apart from this, they have shown their significant contribution in the development of corpus and statistical dictionaries for English-Hindi.

2.12 English-Hindi Machine Aided Translations (MAT) System

Based on a rule-based approach, this system was developed by Jadavpur University, Kolkata for translating specific domain consisting of news sentences and government circulars. In this system, the input sentence is parsed utilizing Universal Clause Structure Grammar (UCSG) parser which gives the number, type,

drksujatha23@gmail.com

and interrelationships among different clauses in the sentence and the word groups as output [10]. A reasonable target language equivalent is thus created from the bilingual dictionary for each word. The target language sentence is then generated by identifying sequence of the clauses and the word groups in proper linear order, on the basis of various constraints of the target language syntax and grammar. Postediting tool is also accommodated for altering the translated text. MAT System 1.0 had shown about 40–60% of completely automatic, accurate translations [11].

2.13 English to (Hindi, Kannada, Tamil) and Kannada to Tamil Language-Pair EBMT System

An example-based English language to Hindi, Kannada, and Tamil languages as well as Kannada to Tamil translation system was developed under the guidance of Balajapally et al. (2006) where a set of bilingual dictionaries comprising of a sentence–phrase–word dictionaries along with phonetic dictionary which comprises of parallel corpora and its mapping is utilized for a corpus size of almost 75,000 sentence pairs [1, 3].

2.14 English-Hindi EBMT System

The Department of Mathematics at IIT, Delhi developed as an example-based English-Hindi MTS which uses divergence algorithms for identifying the discrepancy as well as a systematic method for information retrieval of example based [3].

2.15 A Hybrid Approach to EBMT for English to Indian Languages

In 2007, Vamshi Ambati and U Rohini proposed a hybrid approach for English to other Indian languages while making use of EBMT and SMT methods along with minimal linguistic resources (Ambati et al. 2007). Presently, work is going on to develop English-Hindi as well as other Indian language translation systems on the basis of the manual and a statistical dictionary built using an example database consisting of source and target parallel sentences and SMT tools [12].

drksujatha23@gmail.com

2.16 Anuvadaksh

An effort of EILMT consortium [13], this system is based on a hybrid approach that allows translation of sentences from English to six other Indian Languages, viz., Hindi, Marathi, Oriya, Urdu, Bengali, and Tamil [9]. It comprises of the platform as well as technology independent modules, and facilitates the multilingual community, starting from domain-specific expressions of tourism and healthcare and extending into various other domains in a phase-wise manner. The technologies integrated in this system are Tree-Adjoining-Grammar (TAG), Analyze and Generate Rules (Anlagen) and Example-based MT.

2.17 Rule-Based Transliteration System

Chinnakotla M. K, Damani OM P, SatoskarAvijit (2010) have developed the rule-based systems meant for **Hindi to English, English to Hindi, and Persian to English translation.** Character Sequence Modeling (CSM) used on the source side for identifying word origin along with a human-generated non-probabilistic character mapping rule base for creating translation candidates. On the target side, again CSM is used for ranking the generated candidates. The overall efficiency of this system of using Conditional Random Field (CRF) approach of English-Hindi is 67.0%, Hindi to English is 70.7%, and Persian to English is 48.0% is noted.

2.18 Rupantar

It is developed by CDAC, Mumbai (2012) to enabling writing in a given set Indian language using Roman Script. It has the capability of translating text from one language to another and uses a key map based technique for writing and conversion process [14]. This tool is fast, lightweight and easily integrates with other desktop and web applications.

3 Conclusion

This survey described machine translation (MT) techniques in a longitudinal and latitudinal way with an emphasis on the MT development for Indian languages. Additionally, we tried to describe briefly the different existing approaches that have been used to develop MT systems. From the survey, it is found that most of the

drksujatha23@gmail.com

existing Indian language MT projects are based on a statistical and hybrid approach, because Indian languages are morphologically rich infeatures and agglutinative in nature and have encouraged researchers to pick these ways to deal with creating MT frameworks for Indian languages.

References

1. Naskar S, Bandyopadhyay S (2005) Use of machine translation in India: current status. AAMT J 25–31
2. Bharti A, Chaitanya V, Kulkarni AP, Sangal R (1997) Anusaaraka: machine translation in stages. Vivek, A Q Artif Intell 10(3), 22–25. NCST Mumbai
3. Anthony J (2013) Machine translation approaches and survey for Indian languages. Comput Linguist Chin Lang Process 18(1):47–78
4. Sinha RMK, Jain A (2002) AnglaHindi: an English to Hindi machine-aided translation system. In: International conference AMTA (Association of Machine Translation in the Americas)
5. Ananthakrishnan R, Kavitha M, Jayprasad JH, Chandra S, Ritesh S, Sawani B, Sasikumar M (2006) MaTra: a practical approach to fully-automatic indicative English-Hindi machine translation. In: Proceedings of the first national symposium on modelling and shallow parsing of Indian languages (MSPIL-06) organized by IIT BOmbay, 202.141.152.9/clir/papers/matra_mspil06.pdf
6. http://www.cdac.in/html/aai/mantra.asp
7. http://ebmt.serc.iisc.ernet.in/mt/login.html
8. http://shakti.iiit.net
9. Bandyopadhyay S (2004) ANUBAAD—The translator from English to Indian languages. In: Proceedings of the VIIth state science and technology congress. Calcutta, India, pp 43–51
10. Murthy K (2002) MAT: a machine assisted translation system. In: Proceedings of symposium on translation support system (STRANS-2002), IIT Kanpur, pp 134–139
11. http://cdac.in/index.aspx?id=mc_mat_machine_aided_translation
12. Sinhal RA, Gupta KO (2014) A pure EBMT approach for English to Hindi sentence translation system. I J Modern Educ Comput Sci 7, 1–8. Published Online July 2014 in MECS (http://www.mecs-press.org/)
13. Goyal V, Lehal GS (2009) Advances in machine translation systems. National Open Access J 9. ISSN: 1930-2940, http://www.languageinindia
14. Singh S, Dalal M, Vachhani V, Bhattacharyya P, Damani OP, Hindi generation from interlingua (UNL). Indian Institute of Technology, Bombay (India)
15. http://www.academia.edu/7986160/Machine_Translation_of_Bilingual_Hindi-English_Hinglish_Text
16. http://www.academia.edu/3275565/Lattice_Based_Lexical_Transfer_in_Bengali_Hindi_Machine_TranslationFramework

drksujatha23@gmail.com

Fuzzy-Based Analysis of Information Security Situation



Ashish Srivastava and Pallavi Shrivastava

Abstract As information technology has been a backbone of each sector today, security of information has also been a critical issue for every organization. Once we have decided to secure our organization by various means such as closed circuit camera, antivirus, firewalls, and biometric access controls, question comes before us how can we benchmark our security. Benchmarking of organizational security requires measurement of security situation through value of security situation (VSS) and quantitative model of information security situation (QMISS). Currently, a little amount of research has been done in organizational IT security metrics such as VSS. Most of the security parameters input to VSS can be measured only with inherent uncertainties. For example, "the existence of security policy" often has linguistic answers only (started or in progress). Fuzzy logic is known to better suit for uncertainties. In this paper, we worked on fuzzy-based analysis of QMISS model and VSS. This fuzzy-based analysis of the given model investigates model aspects in detail in the fuzzy environment. It is concluded that QMISS model based VSS computation can be implemented using fuzzy logic as mathematical tool. The analysis is done with help of MATLAB simulation of fuzzy Inference System (FIS).

Keywords Organization information security · Fuzzy · Organizational security benchmarking tool · Value of security situation (VSS) · Quantitative model of information security situation (QMISS)

Proposed QMISS model for security value VSS is as follows: [1].

A. Srivastava (✉)
MSRS, CRL, BEL, Ghaziabad, India
e-mail: ashish_1367@yahoo.co.uk

P. Shrivastava
BrainHaul Technologies, Ghaziabad, India
e-mail: pallavishrivastava2k@gmail.com

© Springer Nature Singapore Pte Ltd. 2019

V. S. Rathore et al. (eds.), *Emerging Trends in Expert Applications and Security*,
Advances in Intelligent Systems and Computing 841,
https://doi.org/10.1007/978-981-13-2285-3_4

drksujatha23@gmail.com

1 Introduction

Nowadays, computer-based information system is pervasive in every sector, whether it is home or business or industry or any other domain. These IT systems are the backbone of business processes. With the spread of information system, chances of cyberattack are also increasing. Therefore, we are bound to secure these IT systems otherwise we can lose millions once IT system is compromised. In worst cases, these cyber attacks on IT system can endanger human lives by triggering nuclear wars at extreme.

Organizations are adopting numerous means of security such as closed circuit camera, antivirus, firewalls, biometric access controls, but the question comes before us how can we benchmark our security [1]. Benchmarking requires computation of VSS.

Very less literature is currently available in this field of VSS and information security modeling [1]. Mohammad Asri et al. described a systems security engineering capability maturity model (SSE-CMM) and other models in his work [2]. Similar situation of literature scarcity can also be seen in fuzzy-based analysis of QMISS model for VSS [1]. Although good amount literature is available on fuzzy logic overview as standalone mathematical technique.

In this paper, the scope of work is fuzzy-based computation of VSS based on QMISS and its analysis. The analysis is done with help of MATLAB simulation of fuzzy inference system (FIS).

QMISS is introduced in Sect. 1, for easy referencing as well as an explanation of further sections. Section 2 describes the basics of the fuzzy logic system. Section 3 of this work describes the fuzzy-based analysis of QMISS model and results obtained in the fuzzy implementation of the model. We conclude the paper in Sect. 4 and seeks further possibilities in this area of security situation modeling, measurement, and forecasting work.

2 Basics of Fuzzy Logic

Nowadays, a variety of fuzzy logic applications has increased in number. Recently, it has been a major area of research in real-world issues. The applications range from household items to industrial process control, to automotive applications, and to artificial intelligence systems.

Fuzzy logic is all about the relative importance of precision: How important is it to be exactly right when a rough answer will do? And it is very close to the thinking pattern of a human brain. Fuzzy logic approach to problems solving imitates how a person would take decisions much faster [3, 4]. Fuzzy logic is based on multivalued logic. Instead of dealing with exactness, it works on approximation and imprecise values. Conventional sets take only two values: true or false. But the fuzzy system takes truth value floating between 0 (completely false) and 1 (completely true).

drksujatha23@gmail.com

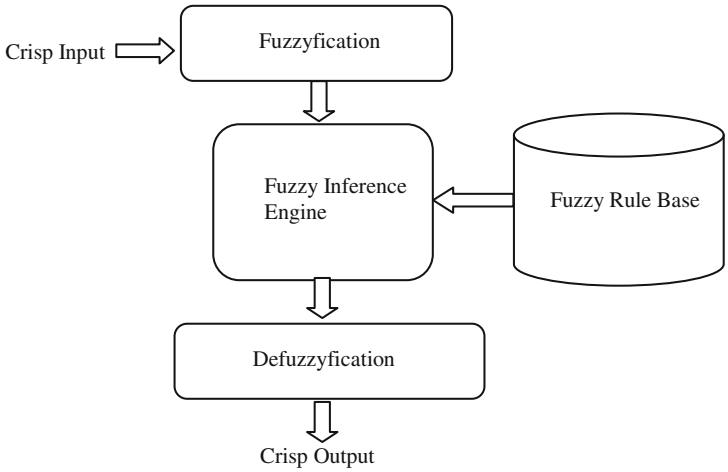


Fig. 1 Block diagram of typical fuzzy logic system

Depending on the value of fuzzy variables also known as linguistic variables, the fuzzy inference engine produces the output values.

The degree of the linguistic variable may be determined with a specific method.

Fuzzy inference systems have a simple concept which majorly includes three steps: Input values, processing the input values and giving output values. The input stage takes crisp inputs values and then fuzzification is done. Fuzzification calculates the degree to which the input values match the fuzzy rule base. Then the fuzzy inference engine calculates the output based on its degree of matching. In the output stage, the results are defuzzified and converted into a crisp value (Fig. 1). Membership function specifies the degree of truth as an extension of valuation. It can be expressed in many forms (Bell curve, triangular, trapezoidal) that specify how each value in the input set is mapped to a membership value (or degree of membership) between 0 and 1.

3 Fuzzy-Based Analysis of QMISS and Discussion

Mamdani fuzzy inference system (FIS) is employed here [refer to Fig. (2)]: FIS model for QMISS). Based on QMISS model (refer Table 1), rules for the FIS system is created. Rules are such as:

- If (ISIs is High) then (VSS is Low) (0.5)
- If (Performance in External Audit is Low) then (VSS is Low) (0.1)

Entire FIS is simulated using MATLAB tools. MATLAB generated surface plots are listed and discussed in Sect. 3.1-"Results and Discussions".

drksujatha23@gmail.com

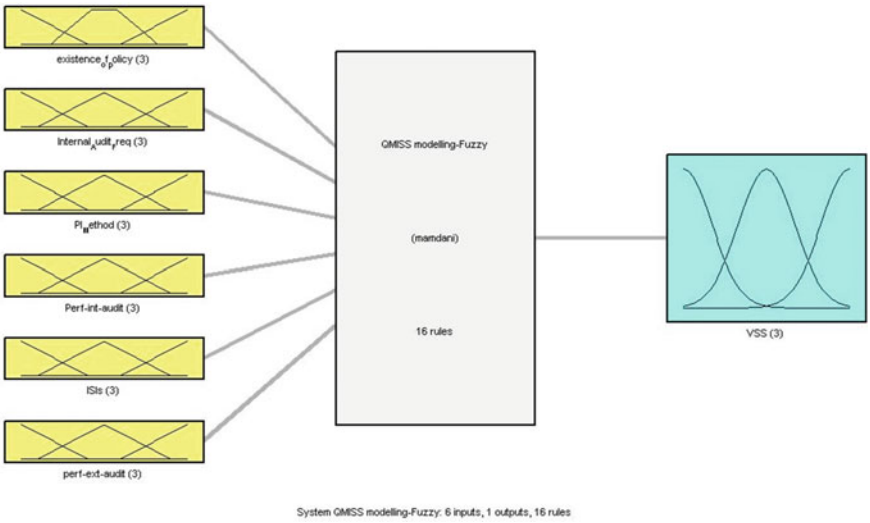


Fig. 2 FIS model for QMISS

Table 1 QMISS parameters

Parameters	Percentage in total value of security situation (%)
Organization's security policy, or an adopted international state of the art policy (P1)	10
Internal audit on Policy (minimum yearly) (P2)	10
Real-time packet inspection (PI) methods (P3)	10
Performance of the users in the Internal audit on Policy (P4)	10
Number of information security Incidents (ISI) reported (P5)	50
External audit performance (P6)	10
Prediction of security situation value as f(Ps) or VSS	Weighted Sum of Above parameters (out of 100%)

3.1 Results and Discussions

In this paper, seven (7) combinations of two inputs against single output "VSS" is plotted in these surface plots.

Discussion for Fig. 3:

Discussion for Fig 3: ISI is 0 then VSS reaches maximum (0.8 in graph). Other input parameters such as PI method are set at 0.5. "Existence of policy" should be at the same time should be maximum at 1.0. While ISI is 0 (best), traversing range 0–1.0 of

drksujatha23@gmail.com

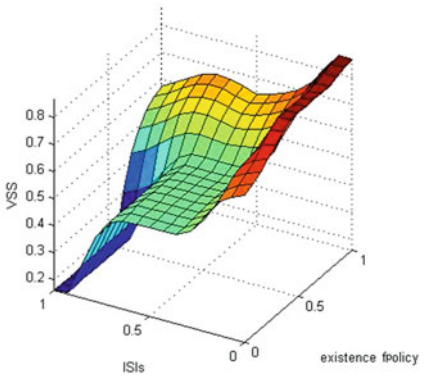


Fig. 3 Surface plot for ISI versus "Existence of policy"

"the existence of policy", generates VSS change from 0.8 to 0.6. While when ISI is 1 (worst), traversing range of existence of policy generates VSS change from 0.6 to almost zero. This is expected inline with QMISS model design as ISI has higher weightage over other parameters of model.

Discussion for Fig. 4:

Both performances in external audit and existence of policy need to be peaked for attaining maximum VSS. While both input parameters are approaching 0 when VSS gets minimizes.

Discussion for Fig. 5:

When we find VSS is peaking out, ISIs value tends to 0 (best) while performance in external audit approached 1 (best),. This performance of Fuzzy Inference system is as per expectation of QMISS. When we analyze ISIs values at 1(worst) within 0–0.5 range of other parameter (performance in external audit) and ISIs values at 0 (best) within 0.5–1.0 range of other parameter (performance in external audit),

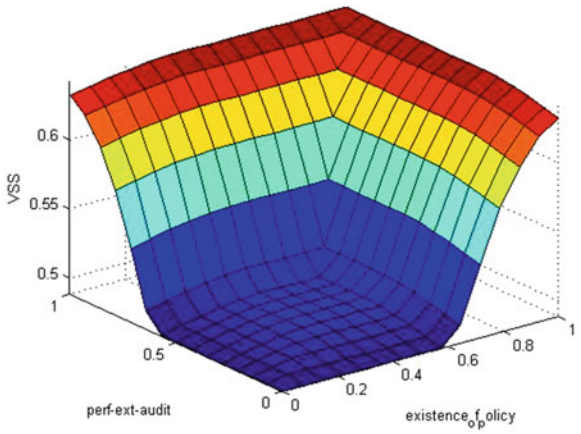


Fig. 4 Performance in external audit (per-ext-audit) versus Existence of policy

drksujatha23@gmail.com

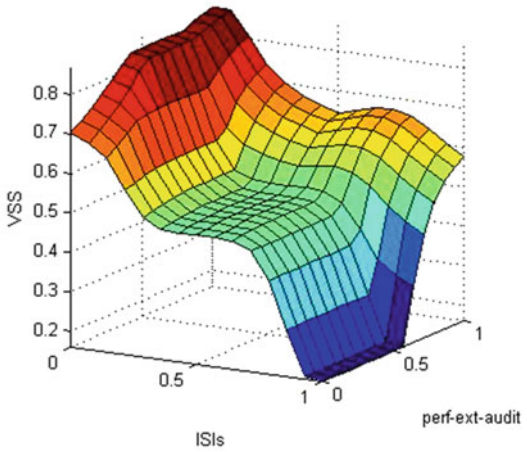


Fig. 5 Surface plot for "Performance in External Audit" versus ISIs

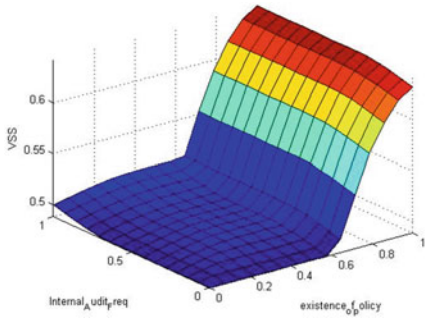


Fig. 6 Surface plot for "Internal Audit Frequency" versus "Existence of policy"

we find VSS has lesser gradient (i.e., more stable). It indicates greater weightage of ISIs performance in overall fuzzy computation of VSS.

Discussion for Fig. 6:

"Internal Audit Frequency" and "Existence of policy" are at worst (0.0) when VSS minimizes. When these both input parameters are at their best (1.0), VSS maximizes, with the higher values of "Existence of policy" which points maturity of policy, VSS gets shows good improvement.

Discussion for Fig. 7:

Here, we see the same result, i.e., VSS maximizes or minimizes with good or bad values of input parameters for packet inspection method or "PI Method" and "Internal Audit Frequency".

Discussion for Fig. 8:

As per expected behavior of FIS, VSS is maximized here once ISIs show the best performance (i.e., ISIs = 0.0). With the improvement in "Performance in Internal Audit", VSS only gets more improved.

drksujatha23@gmail.com

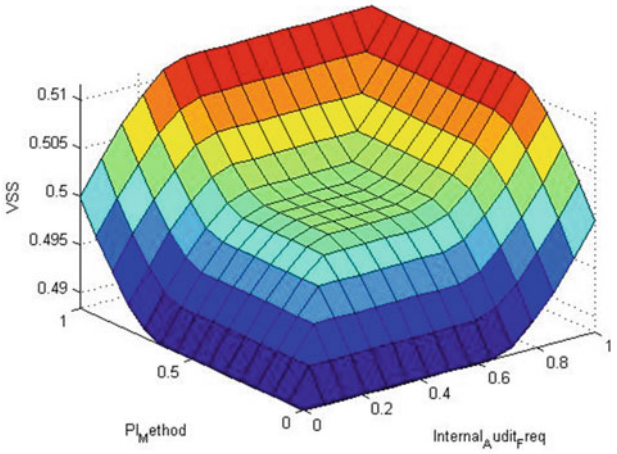


Fig. 7 Surface plot for "Packet Inspection (PI) Method" versus "Internal Audit Frequency"

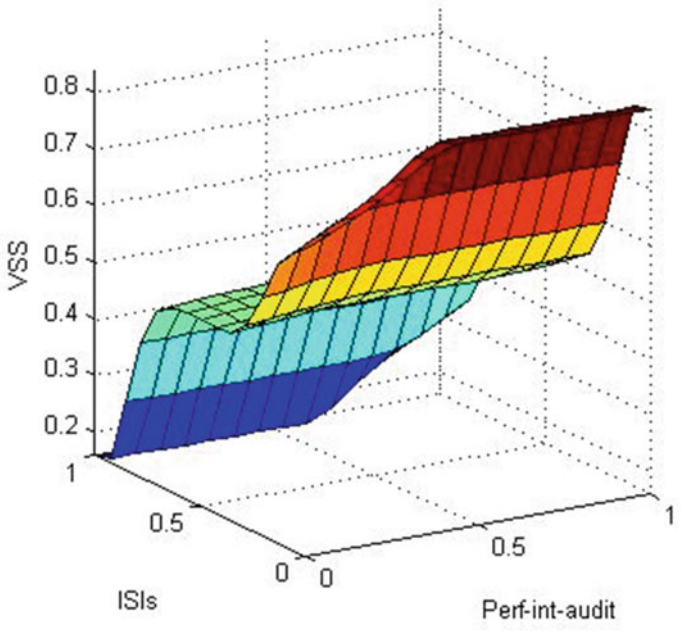


Fig. 8 Surface plot for "ISIs" versus "Performance in Internal Audit (performance-int-audit)"

drksujatha23@gmail.com

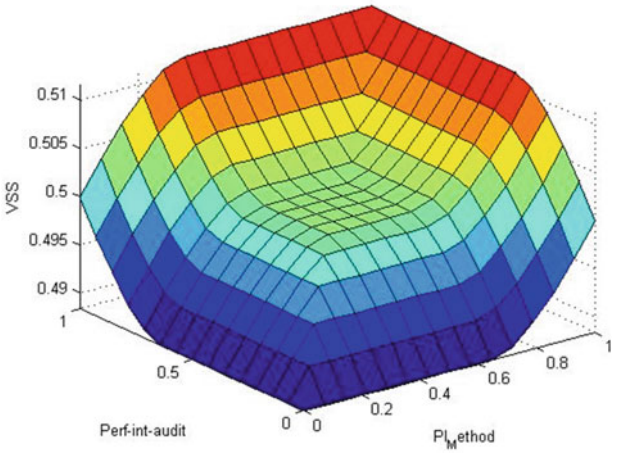


Fig. 9 Surface plot for "PI Method" versus "Performance in Internal Audit"

Discussion for Fig. 9:

This is the same as Fig. 7 where parameters were "PI Method" versus "Internal Audit Frequency". Here, we see same results, i.e., VSS maximizes or minimizes with good or bad values of input parameters for packet inspection method or "PI Method" "Performance in Internal Audit".

4 Conclusion

Imprecise knowledge of constituent parameters of QMISS is very much a possibility given nature of these parameters.

Going with Sect. 3.1 results and discussions, we found that QMISS model based VSS computation can be implemented in non-exact situations using fuzzy logic as mathematical tool. One can evaluate VSS in fuzzy way using QMISS and use VSS for benchmarking, improvement and planning purposes.

Here, in this simulation of Sect. 3.1, we used Mamdani FIS. Sugeno FIS can also be exploited in future works. Artificial intelligence (AI) tools other than fuzzy, such as neuro-fuzzy-based simulation can also be taken up for further work [5–7]. AI tool based simulation of QMISS model is only one group of tools for this broad area of planning, forecasting, measurement, and improvement of information security measurement. So other statistical and mathematical techniques can also be investigated for better measurement and forecasting of information security situation measurement.

drksujatha23@gmail.com

References

1. Srivastava A, Shrivastava P (2018) A study on security situation modelling. Under publication in IEEE
2. Stambul MAM, Razali R (2011) An assessment model of information security implementation levels. In: International conference on electrical engineering and informatics, July 2011
3. Blej M, Azizi M (2016) Comparison of Mamdani-Type and Sugeno-Type fuzzy inference systems for fuzzy real time scheduling. Int J Appl Eng Res 11(2016):11071–11075
4. Nerurkar NW, Kumar A, Shrivastava P (2010) Assessment of reusability in aspect-oriented systems using fuzzy logic. ACM SIGSOFT Softw Eng Notes 35(5), Sept 2010
5. Li C, Hu J, Pieprzyk J, Susilo W (2015) A new bio cryptosystem-oriented security analysis framework and implementation of multibiometric cryptosystems based on decision level fusion. IEEE Trans Inf Forensics Secur 10(6):1193–1206
6. Parra F, Hall LL (2014) A nomological network analysis of research on information security management systems. In: 47th Hawaii international conference on system science, pp 4336–4345
7. Padyab AM, Päivärinta T, Harnesk D (2014) Genre-based assessment of information and knowledge security risks. In: 47th Hawaii international conference on system science, pp 3442–3451

drksujatha23@gmail.com

Estimation of Microwave Dielectric Constant Using Artificial Neural Networks



K. Sujatha, R. S. Ponmagal, G. Saravanan and Nallamilli P. G. Bhavani

Abstract The monotonous and frequent complication in the estimation of dielectric constant expressed interns of frequency in microwave range by incorporating Artificial Neural Networks (ANN). This computerized modus operandi is dependent on the deployment of a slotted line to take measurement which requires a numeric elucidation to resolve the dielectric constant. Automation of the dielectric constant is carried out by developing a computer program for gathering the data acquisition from the conventional setup and modernizing the same using ANN is described here. Investigational data gathered by this existing apparatus is used for training and testing the ANN trained with Back Propagation Algorithm (BPA). An equation formerly obtained from the literature, is used for estimating the dielectric constant. This is compared as an additional function with the computerized algorithm for calibration purpose. Thus, a novel ANN-based scheme for outlining the disparities between various dielectric materials to approximate the dielectric constant is experimentally analyzed using MATLAB.

Keywords Dielectric constant · Back propagation algorithm · Artificial neural networks

K. Sujatha—Masterminded EasyChair and created the first stable version of this document.
Nallamilli P. G. Bhavani—Created the first draft of this document.

K. Sujatha (✉) · R. S. Ponmagal · G. Saravanan · N. P. G. Bhavani
Department of EEE/CSE, Dr. MGR Educational & Research Institute, Chennai, India
e-mail: drksujatha23@gmail.com

R. S. Ponmagal
e-mail: rsponmagal@gmail.com

G. Saravanan
e-mail: saravanang.ece@sairamit.edu.in

K. Sujatha · R. S. Ponmagal · G. Saravanan
Department of ECE, Sri Sai Ram Institute of Technology, Chennai, India

R. S. Ponmagal · G. Saravanan
Department of EEE, Meenakshi College of Engineering, Chennai, India

© Springer Nature Singapore Pte Ltd. 2019

V. S. Rathore et al. (eds.), *Emerging Trends in Expert Applications and Security*,
Advances in Intelligent Systems and Computing 841,
https://doi.org/10.1007/978-981-13-2285-3_5

drksujatha23@gmail.com

1 Introduction

The dielectric constant is defined as the ratio of the permittivity of any material medium to the permittivity of free space. An automation scheme to replace the microwave interferometer has been developed at few Nationalized Experimentation centers for the precise reason of scrutinizing the distress created in permeable materials due to detonation. It is intended that the success of the modus operandi depends on the dielectric constant of the materials under study whose densities are recorded for training purpose. This paper proposes an algorithm, which computes the dielectric parameters and groundwork measurements are done from the existing microwave interferometer system [1]. This ANN-based approach is successful for dielectric constant measurements and uses the existing equipment data for automation. A Personal Computer (PC), used to automate the data acquisition procedure, and a formerly developed MATLAB code for BPA, used to computes the dielectric parameters, were jointly proposed to develop this scheme [2].

2 Existing System for Dielectric Constant Measurement

The system has been automated using a computer for the measurement of dielectric properties, which receives data from a digital meter [3, 4]. This digital meter measures the readings in terms of standing wave ratio, which is compared with the frequency from the microwave source. The reference value of the microwave frequency can be selected automatically using a program run on the computer [5, 6].

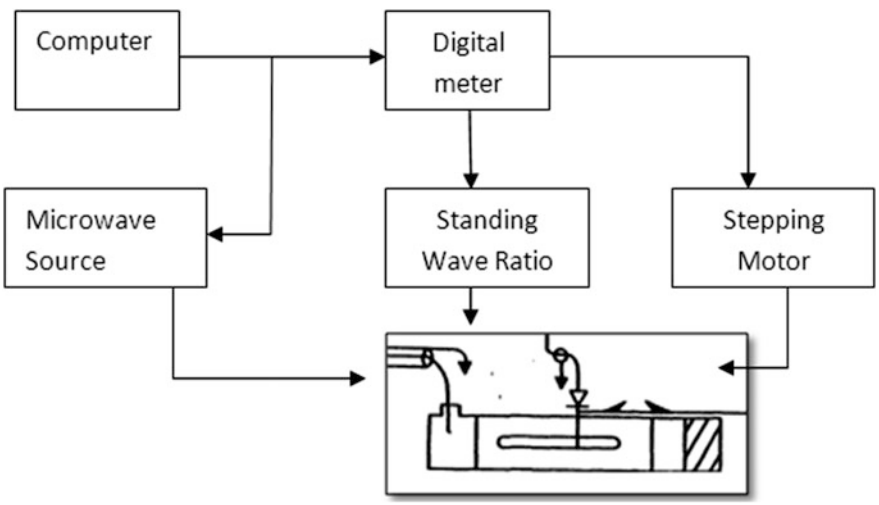


Fig. 1 Existing system for dielectric constant measurement

drksujatha23@gmail.com

Table 1 Input–output data for dielectric constant measurement

Sample	Input parameters				Output Parameter dielectric constant (no units)
	Length (mm)	Density (g/cc)	Temp deg F	Humidity (%)	
Teflon 7C	25.05	0.842	76	35	1.375
Teflon 7C	12.07	1.448	75	46	1.632
Teflon 7C	8.28	2.222	76	44	2.141
Class D HMX	25.05	1.043	76	46	2.148
Class D HMX	12.7	1.387	80	45	2.902
Class D HMX	6.4	1.387	77	50	2.768
Class D HMX	6.4	1.387	78	45	2.807
Class D HMX	6.4	1.387	76	46	2.801
Class D HMX	12.7	1.387	77	46	2.894
Class D HMX	12.7	1.457	81	46	2.87
Class D HMX	11.04	1.599	79	46	3.255
Melamine with no Al	12.06	0.799	77	70	2.259
Melamine with no Al	6.22	0.823	76	69	2.275
Melamine with no Al	12.06	0.823	76	69	2.275
Melamine with no Al	6.22	0.823	76	69	2.276
Melamine with no Al	22.58	0.843	76	66	2.431
Melamine with no Al	12.06	1.053	77	73	2.881
Melamine with no Al	12.06	1.099	77	74	3.055
Melamine with no Al	10.03	1.107	76	67	3.108
Melamine with no Al	9.98	1.121	76	68	3.201
Melamine with 10% Al	6.22	0.74	77	66	2.096
Melamine with 10% Al	12.14	0.76	76	65	2.271
Melamine with 10% Al	12.06	0.842	76	69	2.41
Melamine with 10% Al	6.22	0.851	75	69	2.371
Melamine with 10% Al	6.39	0.984	75	69	2.743
Melamine with 10% Al	9.47	1.113	76	69	3.125

(continued)

drksujatha23@gmail.com

Table 1 (continued)

Sample	Input parameters				Output Parameter dielectric constant (no units)
	Length (mm)	Density (g/cc)	Temp deg F	Humidity (%)	
Melamine with 10% Al	12.14	1.135	77	69	3.18
Melamine with 10% Al	7.29	1.193	76	69	3.404

The principle of operation is shown in Fig. 1. The data collected has been recorded in Table 1.

3 Proposed Automated Dielectric Constant Measurement System

In the proposed system, the estimation of dielectric constant is done using ANN trained with BPA. The data shown in Table 1 is normalized using the formula $X_i/X_{max.}$ The ANN is of feedforward type and uses supervised learning algorithm [7]. The procedure consists of two phases; one the training phase and the other the testing phase. During the training process, the proposed ANN architecture is made to learn the data set for various dielectric samples whose dielectric type, physical properties like length, density, temperature, and humidity serve as input parameters and the target value is the appropriate dielectric constant of the material chosen. The challenge lies in the finalization of the ANN architecture. The block diagram in Fig. 2 represents the dielectric constant automation system.

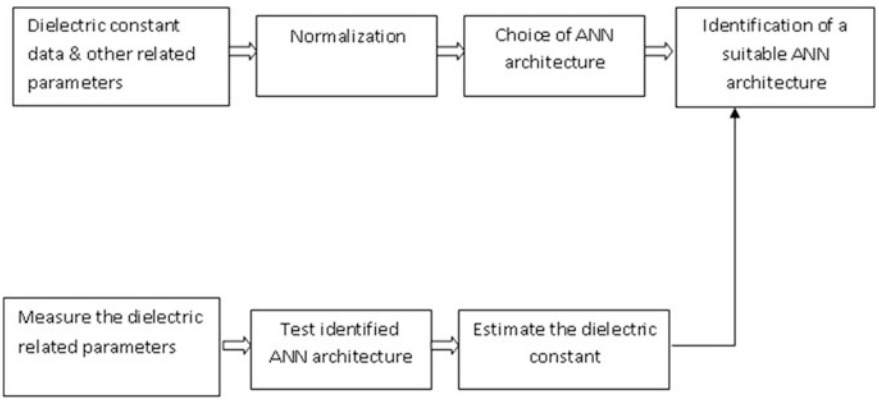


Fig. 2 ANN-based dielectric constant automation system

drksujatha23@gmail.com

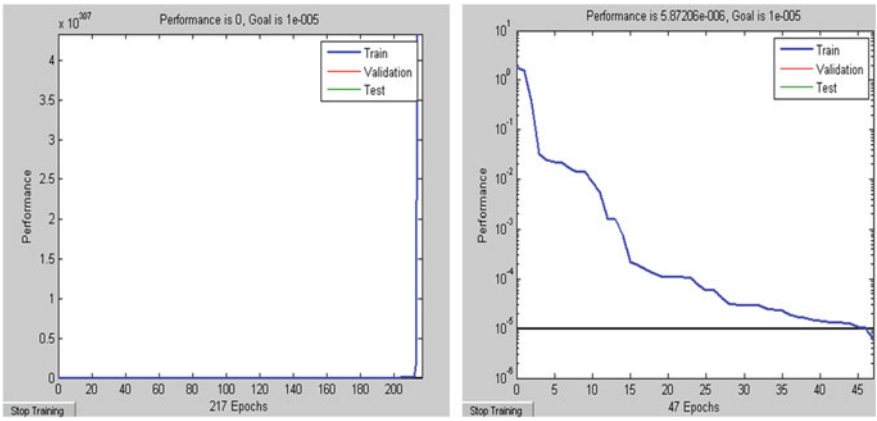
4 Materials and Methods

To implement the proposed ANN-based dielectric constant Automation system a feedforward neural network trained with BPA is used. The feedforward neural network is a multilayer architecture with an input layer, which serves as a buffer, a hidden layer, and an output layer. The layer consists of Processing Elements (PEs) which are capable of performing the computation of inner product. Choice of the appropriate ANN architecture for the automation of dielectric constant is decided based on the optimal value of MSE, number of iterations, and the number of nodes in the hidden layer.

5 Results and Discussion

The routine experimental setup is very tedious and in order to improve the preciseness of measurement of dielectric constant a novel scheme using ANN is proposed here. The initial set of readings can either be collected from the existing setup as in Fig. 1 or else a benchmark equation based on the survey can be used as the input data for training the ANN.

The training of ANN is done using Conjugate Gradient and Quasi-Newton. The results for all the four algorithms are depicted in Fig. 3. The convergence of the conjugate gradient algorithm takes place with minimum number of iterations as well as with the lowest value of Mean Squared Error (MSE). The proposed ANN architecture is also tested and validated using the remaining data set for the automated measurement of dielectric constants.



(a) Conjugate Gradient Algorithm **(b)** Quasi Newton Algorithm

Fig. 3 Training of the feedforward architecture using various types of BPA

drksujatha23@gmail.com

6 Conclusion

This indigenous measurement of dielectric constant using ANN has facilitated the usage of various porous and solid materials to be used in many other practical applications. This method provides a path for matching the impedance precisely.

References

1. Sirikulrat K, Sirikulat N (2008) Dielectric properties of different maturity soybean. KMITL Sci J 8(2):12–18
2. Nelson SO, Wen-chuan G, Samir T, Stanley JK (2007) Dielectric spectroscopy of watermelons for quality sensing. Meas Sci Technol 18:1887–1892
3. Klingensmith JD, Shekhar R, Vince DG (2000) Evaluation of three-dimensional segmentation algorithms for the identification of luminal and medial–adventitial borders in intravascular ultrasound images. IEEE Trans Med Imaging 19(10)
4. Abd-Elmoniem KZ, Youssef ABM, Kadah YM (2009) Real-time speckle reduction and coherence enhancement in ultrasound imaging via nonlinear anisotropic diffusion, vol 6, no 3
5. Perona P, Malik J (1990) Scale-space and edge detection using anisotropic diffusion. IEEE Trans Pattern Anal Mach Intell 12
6. Florack LMJ, Romeny BMH, Koenderink JJ, Viergever MA (2000) Scale and the differential structure of images. Image Vis Comput 10 (1992). Roven Press, New York (1987). In the third trimester, Obstet Gynecol 95(4)
7. Sujatha K, Pappa N (2011) Combustion quality monitoring in PS boilers using discriminant RBF. ISA Trans 2(7):2623–2631

drksujatha23@gmail.com

Bone Fracture Detection from X-Ray Image of Human Fingers Using Image Processing



Anil K. Bharodiya and Atul M. Gonsai

Abstract Orthopaedics deals with surgery and treatment of the human musculoskeletal system. It also involves degenerative conditions, trauma, sports injury, tumors, and congenital issues. Orthopaedic doctors are always interested to take an X-Ray image of injured parts of patient's body for better diagnosis. In an X-Ray imaging, electronic radiation is passed in the human body for capturing bone images. After X-Ray image retrieval, a doctor examines X-Ray image manually. It is not that easy to detect most of the major diseases/issues related with the bones just by visualizing an X-Ray image, although in some cases, it is possible, but till that time, diseases may reached towards next or serious stage for example bone fracture. The main problem with X-Ray images is that they may be blurred, out of focus, improperly bright and noisy, which makes examination more difficult. One of the solutions to all above problems can be computerized image processing of human being's X-Ray images. In this research paper, we have presented an algorithm to detect bone fracture from X-Ray images of human fingers using image processing.

Keywords Orthopaedics · Musculoskeletal · Trauma · X-Ray
Electronic radiation · Image processing

1 Introduction

In a recent medical revolution, computer-aided diseases detection (CADD) plays an important role. The only objective of CADD is to speed up medical diagnosis with greater accuracy and patient's satisfaction. There are many biomedical imaging

A. K. Bharodiya (✉)
UCCC & SPBCBA & SDHG College of BCA & I.T. (BCA Department),
Surat 394210, Gujarat, India
e-mail: anilbharodiya@gmail.com

A. M. Gonsai
Department of Computer Science, Saurashtra University, Rajkot 360005, Gujarat, India
e-mail: atul.gosai@gmail.com

© Springer Nature Singapore Pte Ltd. 2019

V. S. Rathore et al. (eds.), *Emerging Trends in Expert Applications and Security*,
Advances in Intelligent Systems and Computing 841,
https://doi.org/10.1007/978-981-13-2285-3_6

drksujatha23@gmail.com

technologies available like radiography, CT scan, ECG, ultrasound, MRI, etc. In X-Ray imagining the minor amount of electronic radiation is passed through the human body to capture bone images of human anatomy to examine bone-related diseases such as fractures, dislocations, bone degeneration, location of foreign object, infections, osteoporosis, tumors, etc.

Bone injury is a normal issue in human. It happens because of extra load on bone, accident, osteoporosis, bone cancer, heredity, etc. Hence, truthful investigation of bone-related injury is crucial aspect in biomedical analysis. Mainly, radiologists capture and review X-Ray images and finally, design a report with their findings to aid in diagnosis. Further, X-Ray images and report are used by orthopaedics doctors or physicians to prescribe proper treatment to patient. According to paper [1] in X-Ray imaging, a trivial dose of radiation is passed into body to capture images of the interior bone parts of the structure. In the human body, every bone plays an important role for example arm, leg, scalp, etc. Figure 1a (Courtesy: https://images.google.com/) shows X-Ray image of right arm's fingers of human and Fig. 1b (Courtesy: https://images.google.com/) shows X-Ray image of left arm's fingers of human (male/female).

From patient's perspective X-Ray imaging is very better in terms of low cost, first look examination of bone injury, quick response and review, less time required to capture an X-Ray image, etc. From orthopaedics doctor or physicians' perspective, they are always interested to analyze X-Ray image of human being's bone for better diagnosis of the bone related issues. Once the X-Ray image is available, it is manually analyzed by orthopaedics doctors. In the X-Ray image processing, researchers have worked on the computerized analysis of X-Ray images. An automated analysis of X-Ray image may work more concise and powerful for detecting bone related issues such as fracture, osteoporosis, tumors, etc.

The aim of this paper is to present an algorithm to detect bone fracture from X-Ray images of human fingers using image processing. The outcome of this study is to help orthopaedics doctor or physician to identify fracture from bone fingers.

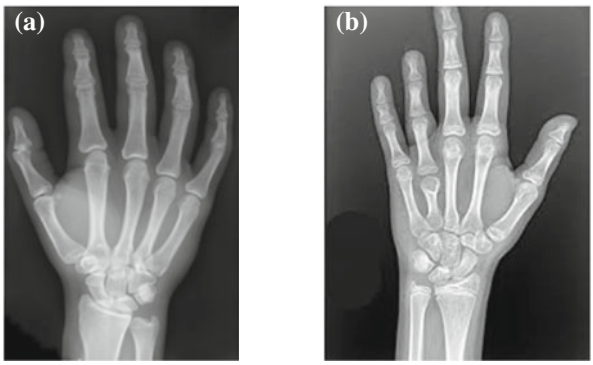


Fig. 1 **a** X-Ray image of right arm's fingers of human, **b** X-Ray image of left arm's fingers of human

drksujatha23@gmail.com

The structure of this article is as follows: Sect. 2 discusses on literature review. Section 3 explains on the proposed algorithm to detect fracture from X-Ray images. Section. 4 discusses analysis of the algorithm. The conclusion is in Sect. 5.

2 Related Works

Major application of image processing that we have perceived in the modern days is in the field of biometric and biomedical image processing. It is nothing but to process on images based on specific algorithms. Medical imaging process is the technique of presenting visually the interior of a body for clinical diagnosis. In modern days, various types of medical imaging technologies are available to do better diagnosis of the human body.

Bones custom the human body structure and support in movement in ground or field. A common bone disorder is the fracture. Fracture is a type of bone condition in which there is a crack due to break down of basic anatomy of bones. Bone crack may happen in many ways. There are three normal situations such as accidents [2], bone malignancy [3], and heavy load [4]. The following paragraphs discuss the work carried out by many researchers to detect fracture from X-Ray images using image processing algorithms.

Umadevi, N. et al. [5] have presented a paper on bone anatomy and fractures. This paper discusses functions and components of human skeleton system such as shapes, region, and structures of human bones; and also describes various types of human bone fractures such as closed (simple), open (compound), comminuted, transverse, blique, spiral, and greenstick (incomplete).

Anu, T. C. et al. [6] have worked on bone fracture detection using image processing. They have proposed a method of fracture detection in steps such as X-Ray bone image collection, preprocessing, edge detection, segmentation, feature extraction, and classification. Accuracy of this method is up to 85%. Nascimento, L. et al. [7] proposed a method to detect fracture from ultrasound image of human being. The proposed method includes steps such as noise reduction, bone line identification, and detection.

Paper [8] explains about the 3D ultrasound image processing such as heart cardiology, obstetrics detection, sonography, and robotics mechanism to detect diseases. Authors of this paper have used the morphological gradient technique to do basic processing of digital images.

Swathika, B. et al. [9] have shown that bone crack can be easily detected with the combination of canny and morphological gradient technique.

Umadevi, N. et al. [10] have worked on classifier to detect fracture from X-Ray. They have worked on a new algorithm which includes steps such as preprocessing, segmentation, feature extraction and normalization, train, evaluation, and mixer of classification. They have used three binary classifier methods such as BPNN, SVM, and KNN. Finally, it was concluded that the combination model of texture and

drksujatha23@gmail.com

shape with BPNN + SVM + KNN results in critical growth in certainty, rigor, time saving, etc.

Zheng Wei et al. [11] proposed an algorithm for the extraction of features from X-Ray images having fracture. This includes steps such as read an X-Ray image, convert into single pixel image, boundary tracking, marker processing (region number, area & centroid), draw protuberant polygon, perform hough transform, detect lines, compute the slope of the line and finally, compute the angle.

Authors of paper [12] have developed a novel algorithm to detect fracture from X-Ray using stacked random forest method. Paper [13] discussed on an algorithm to detect hand bone fracture using image preprocessing, feature extraction, and selection and interpretation. Authors have used MATLAB and WEKA.

Paper [14] proposed a method to extract bone from human X-Ray images. This method includes seven steps such as accept input, enhancement of input using a guided filter, edge detection using canny method, image boundary formation, bone boundary formation and finally, interpretation.

Irfan Khatik [15] has conducted a study of various bone fracture detection techniques. He has discussed on active contour model, Wavelet and Curvelet and Haar, Support Vector Machine (SVM) classifier, X-Ray/CT auto-classification of fracture (GLCM), gradient-based edge detection technique, etc. This paper also described discussion image processing tools such as MATLAB and OpenCV using Python.

It is much clear from the above-related work that many researchers are working in the field of CADD. We are inspired from the above extensive work and propose a novel algorithm to detect bone fracture from X-Ray image of human fingers.

3 Proposed Method

Digital image processing is a wide area and medical image analysis is one of the applications of it. Here, we propose an algorithm to detect bone fracture from X-Ray of human fingers. The proposed algorithm includes systematic steps to detect fracture. The algorithm given in Fig. 2 can be developed in Scilab, Matlab, Octav, OpenCV using Python, etc.

The above Fig. 2 shows the proposed algorithm which includes seven steps to detect fracture from X-Ray image of human fingers. In the first step, X-Ray image of human fingers is inputted into the system. Second step processes the inputted X-Ray image using preprocessing. This step is consists of two sub-steps. In the first sub-step preprocessing is performed using CLHE (Contrast Limited adaptive Histogram Equalizer). According to Paper [16], it modifies the gray-level value based on some characteristics to improve the contrast and minimize the noise. In the second sub-step, guided filter is applied on X-Ray image for smoothening. According to papers [14, 17] in guided filter, a guidance image is used for the noise removal purpose. While fast, this method also preserves edges and gradients and smoothens out the other parts. Third step detects edge of the X-Ray image. As per

drksujatha23@gmail.com

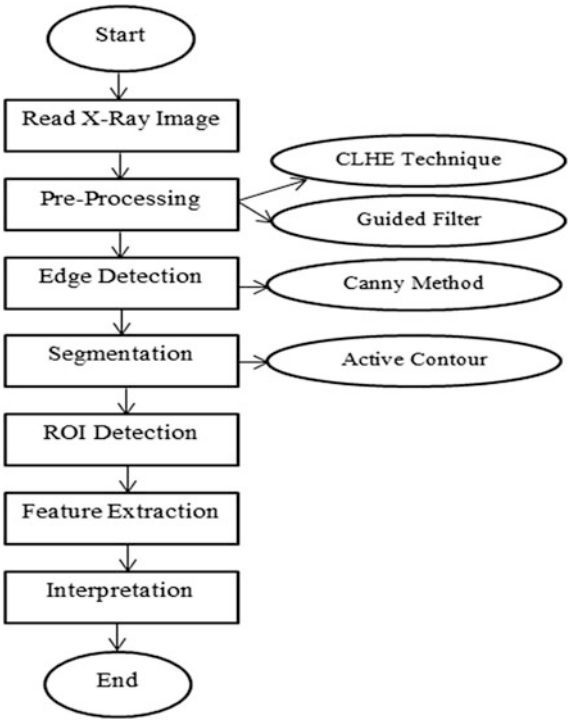


Fig. 2 Algorithm to detect bone fracture from X-Ray image of human fingers

Paper [18, 19] canny is the best method for detection of edges in the image. Further, in the fourth step, X-Ray image is segmented into different regions based on active contour segmentation method. As per Paper [20], active contour method for image segmentation is highly recommended. In step 5, all the segments obtained in the last step are searched to find out ROI. The main purpose of this step is to find out segment of X-Ray image where the region of interest (bone fracture) tends to be available. Further, in the sixth step, features of bone fracture is to be identified from region of interest such as region area, region number, subregion description, magnitude of the fracture edge, direction of the fracture edge, the angle between fracture line and perpendicular line. Finally, in the last step, the interpretation of the bone fracture is to be carried out based on features extracted in the last step.

4 Evaluation and Result

For evaluation of the above algorithm, we want to consider a dataset of 50 X-Ray images, with or without fracture. Images are collected from hospitals in person or downloaded from internet. Since images are in different resolution/size, they are

drksujatha23@gmail.com

fixed in specific size say 512 * 512 or 600 * 700 after enhancement. We proposed to apply above algorithm in Scilab image processing tool (open source) and expect more than 80% accuracy.

5 Conclusions

This paper proposes an algorithm to detect bone fracture from X-Ray image of human fingers. The algorithm is explained in Sect. 3. Compared with previous work, this paper explores bone fracture detection only from human fingers. The above algorithm can be expanded by covering other parts of the human body.

References

1. Bhowmik M, Ghoshal D, Bhowmik S (2015) Automated medical image analyser. In: IEEE ICCSP 2015 conference 0974-0978
2. Rogers LF, Talianovic MS et al (2008) Diagnostic radiology: a textbook of medical imaging, 5th edn. Churchill Livingstone, New York
3. Lee NK, Sowa H, Hinoi E et al (2007) Endocrine regulation of energy metabolism by the skeleton cell. NIH Public Access 130(3):456–469
4. Tamisiea DF (2008) Radiologic aspects of orthopedic diseases. In: Mercier LR (ed) Practical orthopedics, 6th edn. Mosby Elsevier, Philadelphia, PA
5. Umadevi N, Geethalakshmi SN (2011) A brief study on human bone anatomy and bone fractures. IJCES Int J Comput Eng Sci 1(3):93–104
6. Anu TC, Raman R (2015) Detection of bone fracture using image processing methods. Int J Comput Appl 6–9
7. Nascimento L, Ruano MC (2015) Computer-aided bone fracture identification based on ultrasound images. In: IEEE 4th Portuguese bio engineering meeting
8. Singh V, Elamvazuthia W et al (2015) Modeling of interpolation methods for robot assisted and freehand ultrasound imaging system. Procedia Comput Sci 76:15–20
9. Swathika B, Anandhanarayanan K et al (2015) Radius bone fracture detection using morphological gradient based image segmentation technique. Int J Comput Sci Inf Technol 6 (2):1616–1619
10. Umadevi N, Geethalakshmi SN (2012) Multiple classification system for fracture detection in human bone X-Ray images. IEEE ICCNT
11. Zheng W et al (2009) Feature extraction of X-Ray fracture image and fracture classification. In: IEEE international conference on artificial and computational intelligence, pp 408–412
12. CaoY, Wang H et al (2015) Fracture detection in X-Ray images through stacked random forests feature fusion. In: IEEE 12th international symposium on biomedical imaging, pp 801–805
13. Hmeidi I, Al-Ayyoub M et al (2013) Detecting hand bone fractures in X-Ray images. In: Proceedings of world congress on multimedia and computer science organized by ACEEE
14. Kazeminia S, Karimi1 N et al (2015) Bone extraction in X-Ray images by analysis of line fluctuations. In: IEEE international conference on image processing (ICIP), pp 882–889 (2015)
15. Khatik I (2017) A study of various bone fracture detection techniques. Int J Eng Comput Sci 6 (5):21418–21423

drksujatha23@gmail.com

16. Ikhsan IM, Hussain A et al (2014) An analysis of X-Ray image enhancement methods for vertebral bone segmentation. In: IEEE 10th international colloquium on signal processing and its applications, pp 208–211
17. He K, Sun J, Tang X (2013) Guided image filtering. IEEE Trans Pattern Anal Mach Intell 35(6):1397–1409
18. Satange DN, Chaudhari KK et al (2013) Study and analysis of enhancement and edge detection method for human bone fracture X-Ray image. Int J Eng Res Technol (IJERT) 2(4):1196–1202
19. Roopa H, Asha T (2016) Segmentation of X-Ray image using city block distance measure. In: IEEE international conference on control, instrumentation, communication and computational technologies, pp 186–189
20. Mohammadi MH, Guise J (2017) Enhanced X-Ray image segmentation method using prior shape. IET Comput Vis 11(2):145–152

drksujatha23@gmail.com

Review of Data Analysis Framework for Variety of Big Data



Yojna Arora and Dinesh Goyal

Abstract Big Data is too large to be handled by traditional methods for analysis. It is a new ubiquitous term, which describes huge amount of data. Dealing with "Variety", one of the five characteristics of Big Data is a great challenge. Variety means a range of formats such as structured tables, semi-structured log files, and unstructured text, audio, and video data. Every format of data has its unique framework for analyzing it. In this paper, we present a detailed study about various frameworks for analyzing structured, semi-structured, and unstructured data individually. In addition, some frameworks, which deal with all the three formats together, are also explained.

Keywords Big Data · Hadoop · Unstructured data · Analytics
Integrated framework

1 Introduction

Big Data can be defined as large amount of data, which requires new technologies and architectures to be able to extract data from it in an interpreted manner as explained by Sagiroglu and Sinang [1]. There has been an exponential growth in it. It includes terabytes to exabytes of data coming from various sources such as transactions on various portals, results of scientific experiments, weblogs, event details, post and updates of social media, and sensor records of various machines as mentioned by Manyika et al. [2] and Laney [3]. Big Data is usually described in

Y. Arora (✉) · D. Goyal
Department of Computer Science & Engineering, Gyan Vihar School of Engineering & Technology, Suresh Gyan Vihar University, Jaipur, Rajasthan, India
e-mail: Yojana183@gmail.com

D. Goyal
e-mail: dinesh8dg@gmail.com

© Springer Nature Singapore Pte Ltd. 2019

V. S. Rathore et al. (eds.), *Emerging Trends in Expert Applications and Security*, Advances in Intelligent Systems and Computing 841,
https://doi.org/10.1007/978-981-13-2285-3_7

drksujatha23@gmail.com

terms of 3 V's: Volume, Variety, and Velocity. Volume refers to the amount of data, velocity is streaming of data at unprecedented rate, and variety deals with various formats of data. These characteristics as given by Gruska and Martin [4] are responsible for new opportunities and challenges to be faced for Big Data analysis, as mentioned in Manyika et al. [2] and Marx [5]. Among all the challenges, Big Data analysis, Big Data storage, and Management and Big Data processing are the major ones as given in Manyika et al. [2] and Dumbill [6]. Big Data processing challenge is addressed in some of the frameworks explained by Fadnavis et al. [7] and Park et al. [8]. Arora and Goyal [9] and Gruska and Martin [4] explain a detailed literature review about various Big Data processing frameworks and analytic methods. The remaining paper is organized as follows: Sect. 2 explains the framework for analysis of structured data. Section 3 explains about analysis of semi-structured data. Section 4 explains unstructured data analysis, Sect. 5 explains the integrated framework, and Sect. 6 concludes the paper.

2 Structured Data

2.1 RDBMS (Relational Database Management System)

Big Data has been becoming a very important constituent in the way organizations are leveraging high-volume data at the right speed to solve explicit data problems. Data warehouse architecture based on traditional RDBMS method is the first way to handle structured Big Data. Implementation of standard data analysis tasks such as selection and joining is significantly faster in RDBMS. It is very cost effective and helpful in the situations where data changes at a slow rate. The generation of huge amount of data at a very high rate gave rise to parallelized RDBMS explained by Pavlo et al. [10].

Data is stored on multiple machines, tables are partitioned on nodes, and the application layer allows access to data on multiple nodes. The main problem with RDBMS is the large increase in data volume. RDBMS finds it difficult to handle such a large amount of data. To solve this problem, RDBMS adds more central processing unit (or CPU) or more memory for database management system to upgrade vertically. Second, most of the data come from semi-structured or unstructured formats, such as from social media, e-commerce, audio, video, text, and e-mail. Relational databases cannot handle unstructured data. In addition, "Big Data" is produced at a very high speed. The RDBMS lacks high speed because it aims to maintain stable data rather than rapid growth. Therefore, relational databases cannot handle "Big Data" that leads to new technologies.

drksujatha23@gmail.com

2.2 Data Mining and Knowledge Discovery

Data mining is the process of extracting useful information or finding out hidden relationship among data. Data mining has come across several stages. Initially it followed the approach of single algorithm on single machine, followed by database with multiple algorithms in second stage. The third stage was in collaboration with Grid Computing. In the next stage, data mining algorithms were distributed. Lastly, parallel data mining algorithms were implemented to work upon Big Data and Cloud Services. The parallel data mining approach is divided into four phases Association Rule Mining, Classification, Clustering and Prediction.

3 Semi-structured Data

Semi-structured data is a form of data which cannot be queried as it does not have a proper structure which confers to any data model. However, it is a mix of formatted and free text fields and it is often self-describing. In order to query semi-structured data, user should have knowledge about the schema. Some of the sources of semi-structured data are revision control data, log files, test reports, etc. Many frameworks have been implemented in past in order to deal with semi-structured data. Some generic tools were implemented using graph as a structure as explained by Cubrannic et al. [11], Fritz and Murphy [12]. Concept-based techniques have also been applied in order to analyze Software Engineering data as mentioned by Deligiannidis et al. [13].

3.1 Concept-Based Approach

In this approach, semi-structured data is analyzed using formal concept lattice as a universal data structure and tag cloud as an intuitive interface to support data exploration. Concepts are made up of objects and attributes. Lattice is constructed directly from context table. After the construction of concept lattice, interface is constructed using Tag cloud (implemented by Hernandiz [14]) as explained in Fig. 1. The last phase is navigation. It is a stepwise process, and it is updated at each step. If the selected set of keywords is the query which can fetch the result, then the extent of focus concept in lattice is query result.



Fig. 1 Semi-structured data analysis [26]

drksujatha23@gmail.com

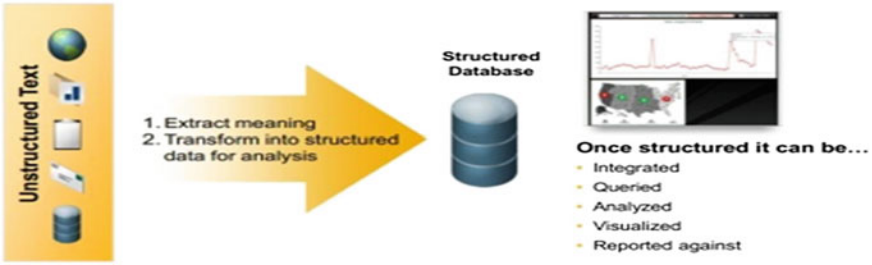


Fig. 2 Unstructured data analysis [16]

4 Unstructured Data

Unstructured data is heterogeneous and variable in nature such as text, audio, video, and images given by Gandomi and Haider [15]. With the digitization of data, unstructured data is growing at a very fast rate. It does not follow any specified format. Unstructured data constitutes 80% of total available data. Unstructured data first needs to be converted into structured data before analysis as explained by Bansal [16] (Fig. 2).

Unstructured data is either machine-generated or human-generated. Some examples of machine-generated data are satellite images, scientific data, and radar or sensor data. Example of human-generated data includes text internal to organization, social media data, mobile data, and website content.

5 Integrated Framework for Structured, Semi-structured, and Unstructured Data

Integrating Big Data aims to build an integrated model for presentation, storage and management, computing, and visualization. It helps to dig deeper into the intrinsic features and characteristics of Big Data.

The integration notation unifies SSU (structured, semi-structured, and unstructured) data and maps Big Data to the same space. The synthesis is represented using the tensor given by Wu et al. [17] and the graph-based structure given by Robinson et al. [18].

Integrated storage and management stores Big Data in a unified storage architecture designed for efficient management, such as faster insertion, deletion, updating, and querying of Big Data. The storage model for unified storage and management is the key value storage model and document database model mentioned by Liu et al. [19].

Big Data Integration: Computing seeks a deeper approach to unifying heterogeneous data. Some applications of depth calculation methods are unsupervised greedy learning algorithm, and Gu and Lin [20] explained convolutional neural networks and stacked automatic encoders.

drksujatha23@gmail.com

Table 1 Comparison of various frameworks implemented for Big Data analysis

S No	Aim	Technique applied/ explained	Type of dataset used	Key features	Findings/results
Unstructured data					
1.	To analyze unstructured data Das and Mohan [23]	Text mining and web mining	Real-time user Tweets and XML files	Data is stored in No SQL HBase for analysis	Unstructured dataset is organized to retrieve knower
2	To analyze unstructured data Lonetey and Deters [24]	Analytics as service tool	NoSQL, HTML, XML, RTF, PDF	Natural language processing	AaaS tool that performs term mining and analytics
3	To develop a framework for real times unstructured Big Data analysis	Complex event processing and continuous query language	CCTV video data taken as object position	Implemented system on wired and wireless network and distributed system environment of RUBA	RUBA framework analyze Big Data using CEP engine and CQL to modify analysis conditions in real time without reexecution of system
4	To analyze unstructured data Vashist and Gupta [25]	Video analytics architecture	Streams of audio and video data	Community Detection, Social Influence Analysis and Link Predication	Video and audio analytics is performed
Semi-structured data					
5	To develop a generic framework for exploration and querying of semi-structured S E data Greene [26]	Concept lattice as universal data structure and Tag cloud for data exploration	XML files and JOSON data	ConceptCloud which creates interactive clouds	A framework for which is fully automated and makes unstructured SE data queriable by tortuously updating keywords tags
6	To visualize semi-structured data Hernandiz and Falconer [14]	Multiple synchronized Tag Cloud	Clinical trial data	Query refinement according to requirement	User can construct an initial query and results are presented in multiple linked tag cloud

(continued)

drksujatha23@gmail.com

Table 1 (continued)

S No	Aim	Technique applied/ explained	Type of dataset used	Key features	Findings/results
Integrated data framework					
8	To Handle Data Heterogeneity Sindhu and Hedge [27]	New algorithms implemented	Patients medical record in text format	Conversion of centralized structured data into distributed structured data, unstructured, and semi-structured to structured data	Proposed system is capable of efficiently performing conversion of centralized to distributed structured data format in less recessing time
9	To implement an integrated Big Data framework Chen and Zhong [28]	Integrated representation Integrated storage Integrated computation Integrated visual analysis	Preprocessing tool for managing Big Data in real time	Graph method Key value method Deep computation	Framework which can deeply mine intrinsic characteristics and features in Big Data
10	To integrate and publish data from multiple sources S. K. Bansal [16]	Semantic Extract Transform and Load (ETL) process	Structured, unstructured, and semi-structured data	Resource description framework: As graph data model	Framework that uses semantic technologies to produce meaningful knowledge
11	To integrate Big Data and Data Mining Reddy and Salim [29]	K-means clustering and Canopy clustering	Structured data	Mahout API working on top of MapReduce	K-means clustering is better for globular data
12	To analyze structured and unstructured data Gharehchopogh and Khalifelu [30]	Text mining and natural language processing	Structured and unstructured data	Web mining NLP algorithms	Structured data obtained using unstructured model of data using text and web mining methods
13	Analysis of Big Data Arora and Chana [31]	Clustering techniques	Real time, Public domain	Portioning generic and density-based algorithms	Current clustering techniques and not efficient for analyzing online streaming and real data

(continued)

drksujatha23@gmail.com

Table 1 (continued)

S No	Aim	Technique applied/ explained	Type of dataset used	Key features	Findings/results
14	To tackle with Big Data's Data variety problem	New framework ePic (scalable and extensible system) based on actor-like programming model	Multi-structured datasets	Concurrent programming model for parallel computations	ePic framework paralyze the programs and runtime system takes care of fault tolerance

Comprehensive Visual Analysis: Due to the lack of completeness, consistency, and accuracy mentioned by Kehrer and Hauser [21], the heterogeneity and dimensionality of the data bring new challenges and opportunities for data visualization. It helps provide multivariable and related presentations of features and associations in Big Data. An et al. have implemented some frameworks to demonstrate interactive visual analysis (2008), Risi et al. [22] and Das and Mohan [23] (Table 1).

6 Conclusion

"Variety" in Big Data has become a great challenge in data analysis. Different approaches need to be adopted for analysis of structured, semi-structured, and unstructured data, respectively. The paper explained some important frameworks, which are implemented for analysis of different formats of data. Structured data analysis can be performed by traditional RDBMS methods and data mining approaches. Semi-structured data on the other hand is analyzed using concept lattice-based approach. Lastly, unstructured data analysis is done by implementing MapReduce methods. An integrated framework can be proposed in future, which can increase the performance of heterogeneous Big Data processing.

References

1. Sagiroglu S, Sinang D (2013) Big data : a review, IEEE
2. Manyika J, Chui M, Brown B, Bughin J, Dobbs R, Roxburgh C, Hung Byers A (2011) Big data: the next frontier for innovation, competition, and productivity. McKinsey Global Institute
3. Laney D (2001) 3D data management: controlling data volume, velocity and variety, In: Application delivery stratergies, Meta Group
4. Gruska N, Martin P (2010) Integrating mapreduce and RDBMS. In: Proceedings of the 2010 conference of the center for advanced studies on collaborative research, pp 212–223
5. Marx V (2013) Biology: the big challenges of big data. Nature 498(7453):255–260
6. Dumbill E (2012) What is big data? an introduction to the big data landscape. Strata
7. Fadnavis RA, Tabhane S (2015) Big data processing using Hadoop. In: IJCSIT, vol 1

drksujatha23@gmail.com

8. Park K, Nguyen MC, Won H (2015) Web based collaborative big data analytics on big data as a service platform. In: ICACT
9. Arora Y, Goyal D (2016) Big data: a review of analytics methods and techniques. In: 2nd international conference on contemporary computing and informatics. IEEE
10. Pavlo A, Paulson E, Rasin A (2009) A comparison of approaches to Large Scale Data Analysis, ACM
11. Cubranic D, Murphy GC, Singer J, Booth KS (2005) Hipikat: a project memory for software development. TSE 31(6):446–465
12. Meyer André, Fritz Thomas, Murphy Gail C, Zimmermann & Thomas, Software Developers' Perceptions of Productivity, In: FSE, Nov 2014
13. Deligiannidis L, Kochut KJ, Sheth AP (2007) RDF data exploration and visualization. In: CIMS, pp 39–46. ACM
14. Hernandiz ME, Falconer SM (2008) Synchronized tag clouds for exploring semi structurd clinical trial data. In: Proceedings of the 2008 conference of the center for advanced studies on collaborative research: meeting of minds. ACM
15. Gandomi A, Haider M (2015) Beyond the hype: big data concepts, methods and analytics. Int J Inf Manage 35(2)
16. Bansal SK (2014) Towards a semantic extract transform load (ETL) framework for big data integration. IEEE
17. Wu X, Zhu X, Wu G, Ding W (2014) Data mining with big data. IEEE Trans Knowl Data Eng 26(1):97–107
18. Robinson EB, Lichtenstein P, Anckarsäter H, Happé F, Ronald A (2013) Examining and interpreting the female protective effect against autistic behavior, Proc Natl Acad Sci USA
19. Liu H, Liu Z, Yuan T, Yao Y (2014) Adaptively incremental dictionary compression method for column-oriented database, pp. 628–632
20. Gu J, Lin Z (2014) Implementation and evaluation of deep neural networks (DNN) on mainstream heterogeneous systems. In: Proceedings of the 5th Asia-Pacific workshop on systems
21. Kehrer J, Hauser H (2013) Visualization and visual analysis of multifaceted scientific data: a survey. IEEE Trans Vis Comput Graph 19(3):495–513
22. Risi M, Sessa MI, Tucci M, Tortora G (2014) CoDe modeling of graph composition for data warehouse report visualization. IEEE Trans Knowl Data Eng 26(3):563–576
23. Das TK, Mohan Kumar P (2013) Big data analytics: a framework for unstructured data analysis. IJET 5(1)
24. Lomotey RK, Deters R (2014) Analytics-as-a-Service (AaaS) tool for unstructured data mining. IEEE
25. Vashisht P, Gupta V (2015) Big data analytics techniques: a survey. IEEE
26. Greene GJ (2015) A generic framework for concept-based exploration of semistructured software engineering data. In: 30th IEEE/ACM international conference on automated software engineering (ASE). IEEE
27. Sindhu CS, Hedge NP (2013) A framework to handle data heterogeneity contextual to medical big data. IEEE
28. Chen Z, Zhong F, Yuan X, Hu Y (2016) Framework of integrated big data: a review. IEEE
29. Reddy V, Arnina Salim MS (2016) A comparative study of various clustering techniques on big data sets using Apache Mahout in 3D. In: MEC international conference on big data smart city. IEEE
30. Gharehchopogh FS, Khalifelu ZA (2011) Analysis and evaluation of unstructured data: text mining versus natural language processing. IEEE
31. Arora S, Chana I (2014) A survey of clustering techniques for big data analytics. In: 5th international conference—confluence the next generation information technology summit. IEEE

drksujatha23@gmail.com

Time Series Forecasting of Gold Prices



Saim Khan and Shweta Bhardwaj

Abstract Data mining is a computing process, for extracting useful information from huge data sets. Using the extracted information, powerful insights such as predictive capabilities and patterns can be acquired. This process of extracting insights from data is also known as KDD (knowledge discovery and data mining). In this paper, time series prediction of gold prices in India is done to predict the price in INR, for gold for the year 2018 up to 31 October 2018. This research paper has used data set for the gold type: MCX Gold, from Quandl. The tool used for modelling this design is RapidMiner, to predict the time series data. This research work has been conducted to predict the price of gold in INR for the year 2018.

Keywords Data mining · Knowledge discovery · Predictive analysis
Time series forecasting · Windowing

1 Introduction

Gold is the most seasoned valuable metal known to man and for a long time, it has been esteemed as worldwide cash, a product, a venture and basically a protest of beauty. In each of the applications it is utilized, gold gives an exceptional execution because of its one of kind properties of being a standout amongst the most pliable metals with high liquefying point and simple recyclability. Gold is a material of decision in pharmaceutical and dentistry as it is biocompatible. As of late, it has been developed as a key nanomaterial. MCX is 'Multi Commodity Exchange'. It was set up in 2003 and is primarily a commodities exchange, providing an online platform for trading, clearing and settlement of commodity derivative instruments.

S. Khan · S. Bhardwaj (✉)
Computer Science and Engineering Department, Amity University Uttar Pradesh, Noida, Uttar Pradesh, India
e-mail: shwetabhardwaj84@gmail.com

S. Khan
e-mail: saimkhan065@gmail.com

© Springer Nature Singapore Pte Ltd. 2019

V. S. Rathore et al. (eds.), *Emerging Trends in Expert Applications and Security*,
Advances in Intelligent Systems and Computing 841,
https://doi.org/10.1007/978-981-13-2285-3_8

drksujatha23@gmail.com

The exchange has a broad national reach with more than 2000 individuals, operations through more than 468,000 exchanging terminals (including CTCL), crossing more than 1900 urban areas and towns in India. MCX is India's driving product fates trade with a piece of the pie of around 81% as far as the estimation of ware contracts exchanged Q1 FY 2014-15.

2 Theoretical Background

The common factors that affect gold prices are as follows:

1. Over the ground supply of gold from national bank deals, recovered scrap and authority gold advances.
2. World macroeconomic factors.
3. Gold demand in India has great seasonal influence.

A. **Data mining**: Information/Data mining is characterized as 'the way toward utilizing an assortment of information examination devices to find examples and connections in information that might be utilized to make legitimate expectations' by 'Two Cross Corporation' (1999). Kumer and Zaki (2000) characterizes information mining as the iterative and intelligent procedure of finding legitimate, novel, helpful and justifiable examples or models in monstrous databases. Information mining is a registering strategy that can be utilized to find obscure examples from the gigantic informational index. The Table 1 given underneath portrays the different factual, NLP, ANN and bunching systems utilized with gold costs information.
There are many existing data mining algorithms/techniques that can be used to extract valuable information from large, complex datasets, which are then analyzed further to discover patterns. All the techniques are broadly classified into two categories:

1. Descriptive: Descriptive techniques are mainly concerned with explanatory models that summarize data for the purpose of interference.
2. Predictive: Predictive techniques are concerned with the creation of predictive models that can produce predictive values when applied to datasets.

Table 1 Details of dataset

Source	https://www.quandl.com/collections/markets/gold
Attributes	Date, Open, High, Low, Close, Interest, Volume
Number of records	209
Name	Gold Futures, October 2017
Time Span	2 January 2017–31 October 2017

drksujatha23@gmail.com

B. **Time Series/Regression**: Time series is utilized for discovering patterns or factual measures and, its analysis is utilized to anticipate future qualities, in light of already existing qualities. Time series analysis can be isolated in two classes: Frequency Domain and Time Domain. Regression is an analytical procedure, utilized for deciding connections between factors. It incorporates different strategies for demonstrating and analyzing of a few factors when the attention is on the connection between a dependant and at least one autonomous factor [1]. Regression analysis is basically utilized as a part of expectation and determining of qualities. This paper employs the use of the tool 'Rapid Miner' with a time series analysis using gold price data.

C. **Time series forecasting using windowing technique in Rapidminer**: Forecasting using RapidMiner, comprises of three steps:

1. Preparing data for Windowing.
2. Model training the dataset [2].
3. Forecasting evaluation [3].

Step 1: **Preparing data for Windowing**: The 'windowing' operator in RapidMiner comes under an extension called 'Series Extension'. The purpose of this operator is to convert series data into generic data. It has various parameters such as 'Horizon', 'Window size' and 'Step size'. **Horizon** gives the upper limit for computation of the prediction. **Window size** gives the number of attributes produced from the data that

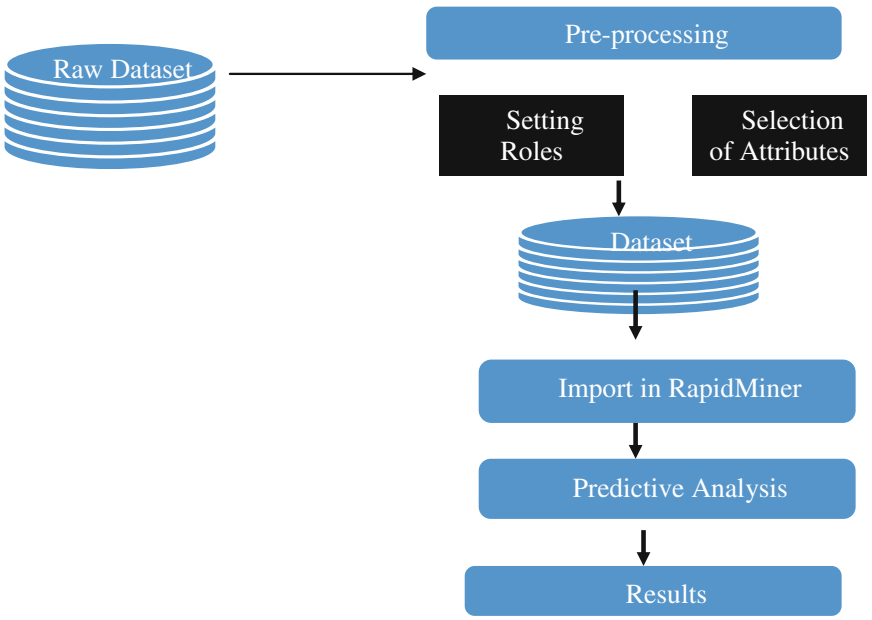


Fig. 1 Methodology

drksujatha23@gmail.com

are cross-sectional. Each record of the first run through arrangement will change over into a new characteristic in the window width. **Step size** decides the technique to move the window.

Step 2: **Model training the dataset**: Sliding Window Validation predictive algorithm is then applied to the dataset, enabling us to further analyze the data using various algorithms. By this step, our time series data has been converted into cross-sectional data, using windowing.

Step 3: **Forecasting evaluation**: We used an 'Apply Model' operator to validate the result generated from the previous model (Fig. 1).

3 Methodology

Step 1: **Raw dataset** consists of multiple attributes like high, low, open, close and date for MCX gold prices. It was collected from Quandl: https://www.quandl.com/collections/markets/gold [4].

Step 2: **Preprocessing** dataset was converted from CSV to Excel Format, useless attributes such as volume and interest were dropped, and missing values were removed. The 'high' attribute, for the highest gold price of the day, was further considered as label and 'date' attribute was used as Id.

Step 3: **Importing Data into RapidMiner** Data is imported in RapidMiner using 'Read Excel' operator. Label is set. Label is the attribute whose value we are predicting. 'High' attribute is used as label [5].

Step 4: **Predictive Analysis** In RapidMiner, time series prediction uses two key data transformations:

1. Windowing transforms the data into a generic set of data, for predicting label attribute that was set in the previous step.
2. Application of the algorithm to forecast target variable and predict time series.

Step 5: **Results** The results produced are validated by Sliding windows validation [6].

drksujatha23@gmail.com

4 Experimental Setup

A. **Dataset**: The dataset used for this research is publicly available from the core financial datasets on Quandl [4]. The present year (2017) data has been utilized for this research. The details of the dataset are given in Fig. 2.

A. **Tool Used—RapidMiner Studio 7.6**: RapidMiner is a software platform, uniting data preparation, machine learning and predictive model deployment under one roof.

The main process applied and explained in the steps mentioned below:

Step 1: **Read Excel**: 'Read Excel' operator reads the data from our excel data sheet and makes the data available in the software platform for further processing [5].

Step 2: **Set Role**: The 'Step role' operator temporarily alters the role of the specified attributes. We alter the role of 'date' attribute to Id [2].

Date	Open	High	Low	Close	Volume	Open Interest
2017-10-31	29692	29698	29490	29616	8584	5078
2017-10-30	29680	29688	29480	29647	5200	4987
2017-10-27	29678	29878	29345	29462	7632	7205
2017-10-26	29343	29434	29302	29549	9044	7374
2017-10-25	29764	29868	29754	29786	7654	6554
2017-10-24	29800	29602	29534	29370	12502	7690
2017-10-23	29460	29876	29071	30032	13290	6833
2017-10-20	29555	29621	29444	29989	1050	816
2017-10-19	29365	29598	29304	30032	10052	1411
2017-10-18	29604	29680	29520	29989	7590	7506
2017-10-17	29410	29550	29522	30032	9044	5711
2017-10-16	29810	29863	29710	29989	7041	7384
2017-10-13	29805	29868	29720	29851	8187	6779
2017-10-12	29775	29900	29710	29814	6812	7107

Fig. 2 Dataset

drksujatha23@gmail.com

Step 3: **Windowing**: The 'windowing' operator is then applied to the dataset, 'High' and 'Low' attributes are selected as labels for prediction. It creates instances of multiple value series utilizing the windowing input data [5].

Step 4: **Validation**: The 'Validation' operator performs cross-validation and statistical performance evaluation of a learning operator that is being evaluated [6].

It comprises of two interrelated processes: 'Training' process is used to train our model, by using a SVM (Support Vector Machine) operator, and then is further applied in the 'Testing' process [7]. Any algorithm obtained from the 'Machine Learning' subfolder can be used for the training process. Other than SVM, neural networks, regression, etc., could possibly be used as well, with appropriate changes in the process design (Refer Fig. 3).

5 Analysis

The gold price prediction was done for the data set ranging from 2 January 2017 to 31 October 2017. The factors under consideration for analysis are:

1. Prediction using high as the label
2. Prediction using low as the label.

Case Study 1: Prediction using High as the label

Actual High (red) and predicted high (blue) are shown on trajectories in Fig. 4. MCX index 2 January 2017 to 31 October 2017 is taken. The trajectories are closely projected and predict high attribute.

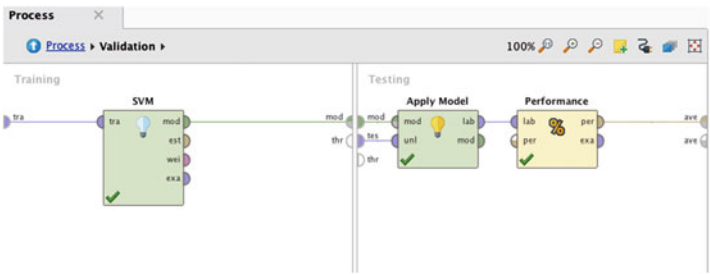


Fig. 3 Validation model

drksujatha23@gmail.com

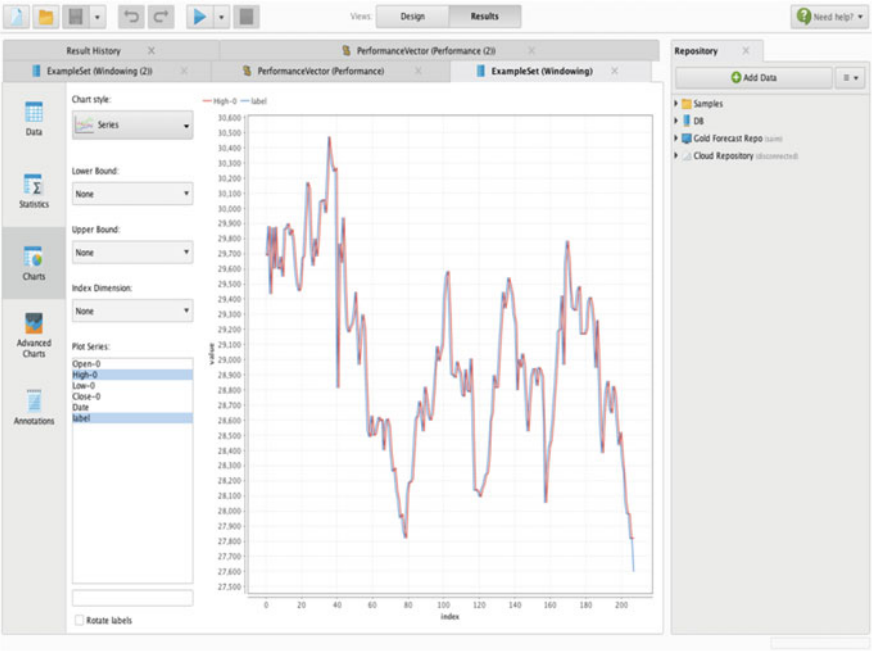


Fig. 4 Time versus actual high versus label

Case Study 2: Prediction using Low as the label

Actual Low (red) and predicted low (blue) are shown on trajectories in Fig. 4. MCX index 2 January 2017 to 31 October 2017 is taken. The trajectories are closely projected and predict the low attribute (Fig. 5).

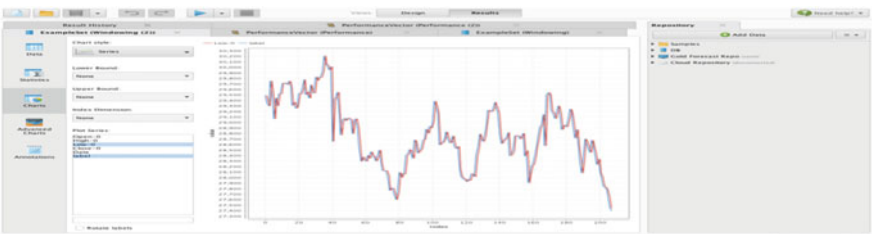


Fig. 5 Time versus actual low versus label

drksujatha23@gmail.com

6 Conclusion

MCX Iindex, complex entities are affected by various components and to entitle that forecasting will anticipate the exact cost of an item is illogical. A bunch of the reasons are MCX index is affected are the same as impacting the gold. In our analysis for Case I, we observed that the highest price for gold in 2017 was INR 30,474, with an average high price of INR 29,0,20.880, while our prediction for 2018 gives the highest price of gold at INR 30,474, with an average high price of INR 29,010.793. In our analysis for Case II, we observed that the lowest price for gold in 2017 was INR 27,450, with an average low price of INR 28,802.125, while our prediction for 2018 gives the lowest price of gold at INR 27,401 and an average low price of INR 28,792.082. In the near future, when smarter algorithms, who adapt even better to learning curves, will be developed, the predicted values will be much closer in comparison to the predictions we can make today. Some weightage can also be given, to the external factors that affect the pricing of any commodity, by estimating to what extent these factors affect the pricing of a commodity, we will be able to predict precise values, and for greater time extent.

References

1. Zainal NA, Mustaffa Z (2016) Developing a gold price predictive analysis using Grey Wolf Optimizer. In: 2016 IEEE student conference on research and development (SCOReD)
2. http://docs.rapidminer.com/studio/operators/blending/attributes/names_and_roles/set_role.html
3. http://docs.rapidminer.com/studio/operators/datatransformation/attribute_space_transformation/selection/selectattributes.html
4. https://www.quandl.com/collections/markets/gold
5. http://docs.rapidminer.com/studio/operators/import/data/read_excel.html
6. http://docs.rapidminer.com/studio/operators/validation/x_validation.html
7. http://www.simafore.com/blog/bid/106430/Using-RapidMiner-for-time-series-forecasting-in-cost-modeling-1-of-2
8. Suranart K, Kiattisin S, Leelasantitham (2014) A analysis of comparisons for forecasting gold price using neural network, radial basis function network and support vector regression. In: The 4th joint international conference on information and communication technology, electronic and electrical engineering (JICTEE)
9. Jie C, Rongda C (2013) Analysis on the impact of the fluctuation of the international gold prices on the gold stocks in Chinese Shanghai and Shenzhen A-Share. In: 2013 sixth international conference on business intelligence and financial engineering
10. Wang J, Kou L, Hou X, Zhou Z (2010) Empirical analysis on co-movement of stock price of gold mine enterprises and the international gold price. In: 2010 international conference on electrical and control engineering
11. Cheung H, Wang M, Lai KK (2013) Stability analysis of influence factors of long-term gold price. In: 2013 sixth international conference on business intelligence and financial engineering
12. Sekar KR, Srinivasan M, Ravidiandran KS, Sethuraman J (2017) Gold price estimation using a multi variable model. In: 2017 international conference on networks and advances in computational technologies (NetACT)

drksujatha23@gmail.com

13. Huang D, Wang S (2010) The real diagnosis analysis of the correlation relations between USD index and gold price. In: 2010 sixth international conference on natural computation
14. https://rapidminer.com/training/videos/
15. Troiano L, Kriplani P (2010) Predicting trend in the next- Conference on, Krackow, pp 199–204

drksujatha23@gmail.com

Titanic Data Analysis by R Data Language for Insights and Correlation



Shaurya Khanna, Shweta Bhardwaj and Anirudh Khurana

Abstract One of the very fatal and tragic events in the history, the Titanic tragedy had an impact on people for about 100 years. During the duration of the Titanic incident, it is believed that the ship charged ahead at speeds higher than what was recommended. The objective of this research paper is to apply different analysis methods of R to dataset to discover the attributes that the surviving passengers possessed. Ggplot2 is also utilized. From the results, the insights are discovered.

Keywords R language · Knowledge discovery · Titanic · Ggplot2

1 Introduction

The Titanic was a ship calamity that on its journey submerged in Atlantic on April 15 1912, killing 1502 passengers and crew. Total number of passengers travelling was 2224. The analysis of survival of passengers is still analysed till this date. The data for Titanic was available at Kaggle repository [1]. Kaggle is a data repository available for public use. This data is interesting to find a correlation between different aspects of the passenger's data.

S. Khanna · S. Bhardwaj (✉) · A. Khurana
Computer Science and Engineering Department, Amity University,
Noida, Uttar Pradesh, India
e-mail: shwetabhardwaj84@gmail.com

S. Khanna
e-mail: shauryakhanna17@gmail.com

A. Khurana
e-mail: khuranaanirudh27@gmail.com

© Springer Nature Singapore Pte Ltd. 2019

V. S. Rathore et al. (eds.), *Emerging Trends in Expert Applications and Security*,
Advances in Intelligent Systems and Computing 841,
https://doi.org/10.1007/978-981-13-2285-3_9

drksujatha23@gmail.com

2 R Language for Big Data Analytics

R is a very diverse language utilized as (mathematical) statistics oriented computational and visualizing tool. R project comprises of two parts, basic part and packages. When we introduce R from the point site of R venture, the R basic venture is set up. R basic part incorporates the methods for easy statistical computing and graphical approach, shown as scientific statistical methods, linear and complex modelling, cluster formation and plotting [2]. But, the limit for advancing in incorporating of supporting vector machinery is evolution of computing, fuzzy clusters, etc. To be able to solve the problem, R provides the packages part. The R packages part includes the ability to utilize R computing and graphical plotting. To incorporate the modules for supporting vector machinery, additionally, we can install the R module part 'e1071' integrated with R basic parts. Module of 'tm' has much functionality for textual miners for features processing and textual data clustering. The module is keenly insightful for analysing of distinct features, because variably big data includes textual datum features. It shows R as systematic and logical tool for feature extraction and scientific incorporation. Two modules present in R system, constitute R to be a complete data science [3]. This combining of packages is utilized for functionality of big data. In this paper, we are focusing upon combination of R systems with functionality of data science. Big Data is the analytical approach of datum. Data science includes every part about datum which is data collecting, transforming, architectural approach, storing, analysing, visualizing and deploying in the environment [4]. So, data science is a disciplinary in need of varied skills in every possible part from life sciences to variety of scientific field, or from mathematical approach to business execution. Transforming and analysing are crucial components in data science fields, because most of datum is without structure, discrete and non-numerical. We shall use novel patterns of the features using analytical approach [2].

3 Why Ggplot2?

A. **Advantages of ggplot2**

- Consistency in overlaying grammar of graphics [5]
- Plotting specified even at higher levels of abstraction
- Flexible
- Themes environment for decorating plotting experience
- Maturity and completion of graphical system
- Multi-users, dynamic mailing lists

drksujatha23@gmail.com

There are some things denied to do with ggplot2:

- Three-dimensional graph plotting (rgl package)
- Graph theory typical graphs (igraph package)
- Interactive plotting and graphics (ggvis package).

B. **What Is The Grammar Of Graphics?**

The basic structure: independently specifying plotting of builder blocks and combining them for creation of any type of graphics display [5]. Building blocks of a graphical system include:

- Datasets.
- Mapping of graphical aesthetic.
- Linear and geometrical blocks.
- Statistical analysis and transforming of values.
- Scaling.
- Coordinated environment.
- Positioning and adjusting.
- Facets.

4 R Program

```
library(ggplot2)
titanic<- read.csv('/Users/Dell/Desktop/train.csv',stringAsFactors=FALSE,header=T)
View(titanic)
titanic$Pclass= as.factor(titanic$Pclass)
titanic$Survived= as.factor(titanic$Survived)
titanic$Sex= as.factor(titanic$Sex)
titanic$Embarked= as.factor(titanic$Embarked)
titanic$Fare= as.factor(titanic$Fare)
titanic$Parch= as.factor(titanic$Parch)
titanic$SibSp= as.factor(titanic$SibSp)

#Survival rate by gender?
#Bar-Graph
ggplot(titanic,aes(x=Sex,fill =Survived)) +
theme_bw() +
geom_bar() +
labs(y="Passenger Count",
title="Titanic Survival Rates by Sex")
```

drksujatha23@gmail.com

```
#Distribution of passenger ages?
  #Histogram is better for visualizing numeric data
ggplot(titanic,aes(x=Age)) +
theme_bw() +
geom_histogram(binwidth = 5) +
labs(y="Passenger Count",
       x="Age(binwidth = 5)",
title = " Titanic Age Distribution")

#Boxplot (Survival rates by age)
ggplot(titanic,aes(x=Survived,y=Age)) +
theme_dark() +
geom_boxplot() +
labs(y="Age",
       x="Survived", title = " Titanic Survival Rates by Age")
#Density-Graph
ggplot(titanic,aes(x=Age,fill=Survived)) +
theme_bw() +
facet_wrap(Sex~ Pclass) +
geom_density(alpha=0.5) +
labs(y="Age",
       x="Survived",
title = " Titanic Survival")
```

5 Results: Correlation Graphs

See Figs. 1, 2, 3, 4, 5 and 6.

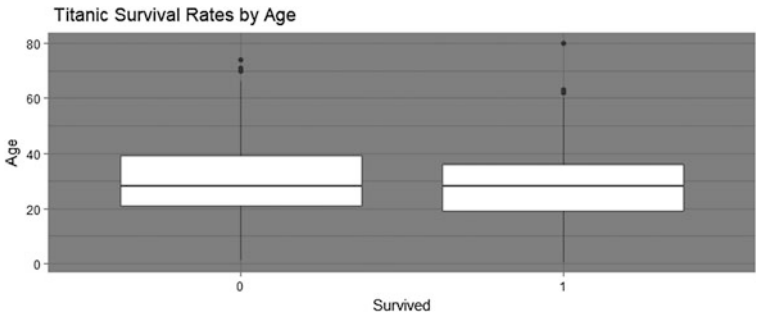


Fig. 1 Titanic survival rates by age

drksujatha23@gmail.com

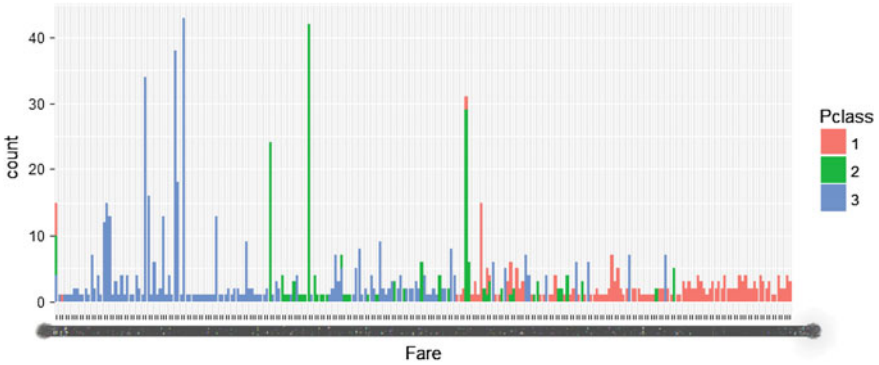


Fig. 2 Court versus fare

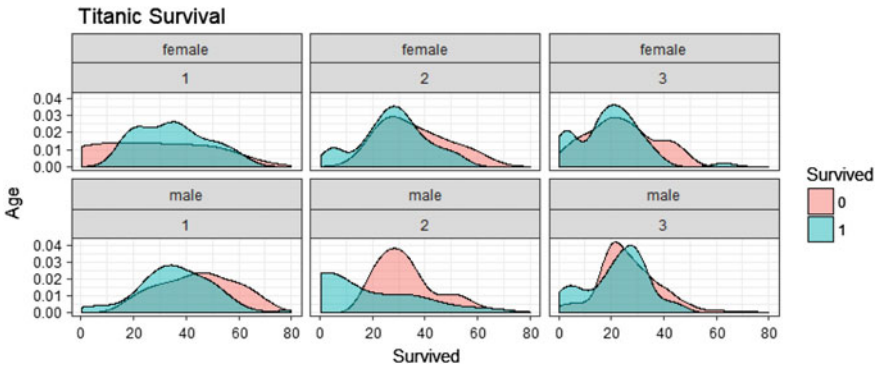


Fig. 3 Titanic survival

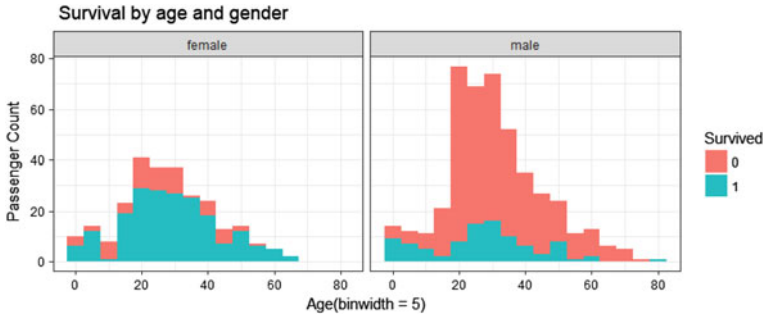


Fig. 4 Survival by age and gender

drksujatha23@gmail.com

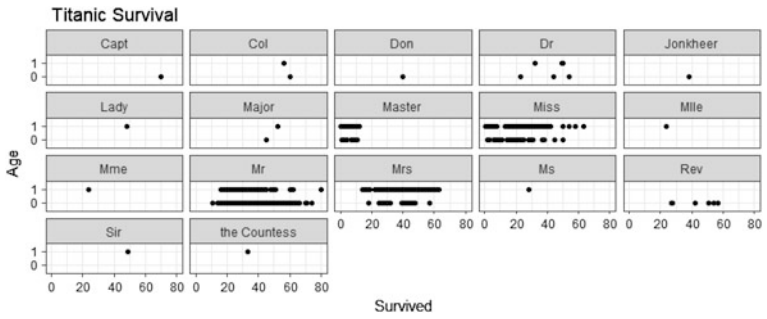


Fig. 5 Age versus survived

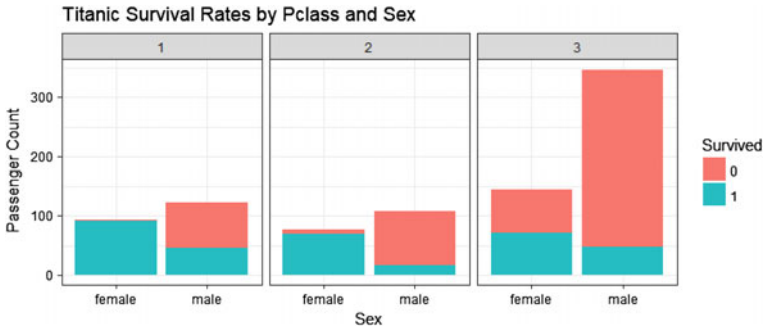


Fig. 6 Titanic survival rates by pclass and sex

6 Analysis

- Only one female of a very young age died who was travelling in Class 1.
- 10% female died from Class 2 and they were above middle age.
- 50% female died from Class 3 and 5% of them were teenagers.
- 65% male died from Class 1 and they were age 40.
- 80% male died from Class 2 and they were between 20 and 60 yrs of age.
- 80% male died from Class 3 excluding ages 1–15 and 25–35.
- 75% of children were able to survive, which included mostly younger girls and older boys.
- Women with title Mrs. had a higher percentage of survival rate.
- 85% people were between age groups 20–40.
- Box plot also shows us that the median age group of people who survived was less than the median age group of people died.

What is proved?

Hypothesis: Women and children survived.

drksujatha23@gmail.com

7 Conclusion

In this specific paper, we are applying R data science for analysing data from Titanic. We mix the R system with data applied science for complete analysis. R is open free software used as statistical communication and visualizing tool. It includes the use of R and other packages able to improve the computable ability of R exponentially. This expresses the power of a complete R system environment. Statistics involves study with big data as well as diverse data. We utilized R as a data language with diverse options. So, the use of R as a data science makes us capable to perform distinct quantitative analysis.

References

1. https://www.kaggle.com/mrisdal/exploring-survival-on-the-titanic
2. Jun S (2016) Patent big data analysis by R data language for technology management. Int J Softw Eng Appl 10(1):69–78
3. R development core team (2015) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org
4. https://www.datacamp.com/community/blog/machine-learning-tutorial-for-r
5. https://github.com/IQSS/workshops/blob/master/R/Rgraphics/Rgraphics.org
6. Chatterjee T (2017) Prediction of survivors in titanic dataset: a comparative study using machine learning algorithms. Int J Emerg Res Manag Technol. Department of Management Studies, NIT Trichy, Tiruchirappalli, Tamilnadu, India
7. Singh A, Saraswat S, Faujdar N (2017) Analyzing Titanic disaster using machine learning algorithms. In: International conference on computing, communication and automation (ICCCA), pp 406–411
8. Biel Steven (1996) Down with the old canoe: a cultural history of the Titanic disaster. W.W. Norton, New York
9. https://www.analyticsvidhya.com/blog/2016/02/complete-tutorial-learn-data-science-scratch/
10. Halpern, S (2011) Report into the loss of the SS Titanic: a centennial reappraisal. Stroud, Gloucestershire U.K., History

drksujatha23@gmail.com

Edge Detection Property of 2D Cellular Automata



Wani Shah Jahan

Abstract Cellular Automata have been used for a wide range of applications. Pattern generation, cryptography, image processing (feature extraction, edge detection, noise cleaning, translation, zooming etc.), and urban growth: prediction and planning, forest, and water body surveying are a few applications to mention here. CA rules have been used for various applications of image processing and I have observed a random selection of rules in most of the studies carried out until now. This study is intended to identify the classes of 2D Cellular Automata linear rules for the quality of edge detection/contour determination of optical and spatial images. I have made an analysis of 2DCA linear rules in Moore neighborhood and classified them according to their property of edge detection and contour determination quality. The results achieved will not only simplify the selection criteria of rules for the purpose but also enhance the precision in recognizing elements and patterns in various types of image data in general and spatial image data in particular.

1 Introduction

Von Neumann and Stanislaw Ulam introduced cellular lattice in the late 1940s as a framework for modeling complex structures capable of self-reproduction [1]. Cellular Automata are based on a concept of dividing space into a regular lattice structure of cells where each cell can take a set of "n" possible values. The value of the cell changes in discrete time steps by the application of rule R that depends on the neighborhood around the cell. The neighborhood can be along a line, in a plane or in space. The neighborhood along a line gives rise to one-dimensional cellular automata (1DCA), neighborhood in the plane is known as two-dimensional cellular

W. S. Jahan (✉)
Department of Electronics, S. P. College, Cluster University,
Srinagar 190001, Jammu and Kashmir, India
e-mail: dr.wani.shahjahan@ieee.org; drsjwani@gmail.com

© Springer Nature Singapore Pte Ltd. 2019

V. S. Rathore et al. (eds.), *Emerging Trends in Expert Applications and Security*,
Advances in Intelligent Systems and Computing 841,
https://doi.org/10.1007/978-981-13-2285-3_10

drksujatha23@gmail.com

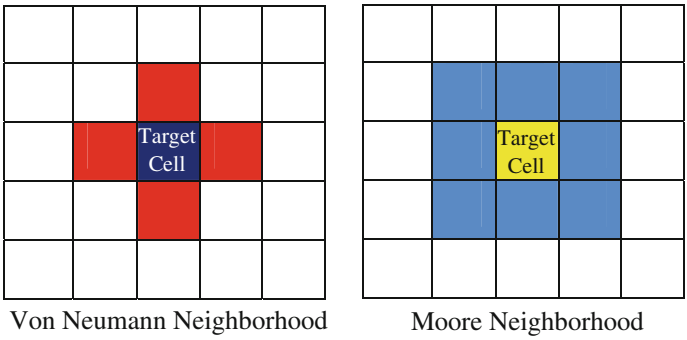


Fig. 1 Commonly used neighborhood structures

automata (2DCA) while as the neighborhood in space is called three-dimensional cellular automata (3DCA).

Cellular Automata (CA) model is composed of a universe of cells in a state having neighborhood and local rule. With the advancement of time in discrete steps, the cell changes its value in accordance to the state of its neighbors in the present state. Thus, the rules of the system are local and uniform. There are one-dimensional, two-dimensional, and three-dimensional CA models. In this study, the 2DCA rules are explored for edge detection property where the central space represents the target cell (cell under consideration) and all spaces around represent its eight nearest neighbors. The structure of the neighbors mostly discussed and applied are Von Neumann and Moore neighborhood shown in Fig. (1).

The two-dimensional Cellular Automata, in general, are represented by the equation as shown below

$$\left[a_{i,j}\right]_{t+1} = R\left[a_{i,j}, a_{i,j+1}, a_{i+1,j}, a_{i,j-1}, a_{i-1,j}\right]_t \tag{1}$$

For linear 2D Cellular Automata, the implementation of the edge detection property investigation the equation can be written as follows:

$$\left[a_{i,j}\right]_{t+1} = \mathrm{XOR}\left[\ a_{i,j}, a_{i,j+1}, a_{i+1,j}, a_{i,j-1}, a_{i-1,j}\right]_t \tag{2}$$

The linear 2DCA attracted a number of researchers for various applications in industry and research. The most important among such applications are the VLSI design, image processing, and graphics. The use of rules for graphical translations has been reported by Qadir, Jahan, Khan, and Peer [2]. Various studies have also been carried out by Pabitra Pal Choudhury et al., who classified the cellular automata rules in Moore neighborhood by assigning the rule values to different cells

drksujatha23@gmail.com

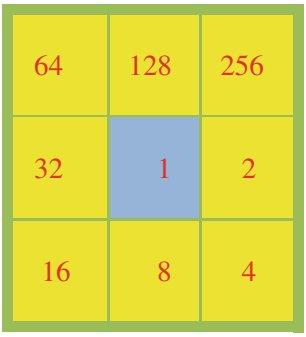


Fig. 2 Rule numbering in linear 2DCA

as shown in Fig. (2). 2DCA linear rules for QCA, evolution, and boundary defects have also been reported by Qadir, Wani et al. [3, 4].

Examples

$$\text{Rule } 3 = \text{Rule } 2 \oplus \text{Rule } 1 \quad \ldots(\text{Group } 2)$$
$$\text{Rule } 11 = \text{Rule } 8 \oplus \text{Rule } 2 \oplus \text{Rule } 1 \quad \ldots(\text{Group } 3)$$
$$\text{Rule } 15 = \text{Rule } 8 \oplus \text{Rule } 4 \oplus \text{Rule } 2 \oplus \text{Rule } 1 \quad \ldots(\text{Group } 4)$$
$$\ldots \text{So on to Rule } 511$$

2 Earlier Classifications

Wolfram's work in classification of CA rules [5] is the most applied and discussed one in the present cellular automata research. He classified the 1DCA rules into four types according to results of their pattern evolution. These classes are Homogenous State Structures, Simple Stable Structures, Chaotic Structures, and Complex Localized Structures. For the two-dimensional classification, Wolfram carried on the pattern of 1DCA rule classification [6, 7]. Li, Packard, and Langton [8] observed phase transition like phenomena between ordered and disordered Cellular Automata dynamics. Ohi [9] observed a chaos in patterns of Rule 40 that again indicates the need for more stable classification of CA rules. Michail, Claude, and Jean [10] reported that the glider rules are located between order and chaos. Choudhary et al. [11], while reporting classification of rules based on their properties has suggested optical devices for implementation. During our experimentation, I have found that the classification on the basis of pattern evolution and edge detection property are more suitable and valid for 2DCA linear rules, where operations are purely Exclusive-OR.

drksujatha23@gmail.com

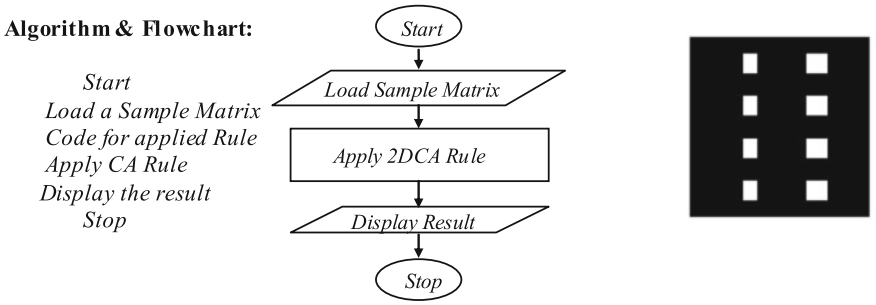


Fig. 3 Algorithm and sample image matrix

3 Proposed Method

Image processing applications of Cellular Automata have made a good mark. Edge detection and contour determination of image data is a widely discussed application of Cellular Automata. During observations, a random selection criteria of CA rules have been reported by the authors. The proposed classification of CA rules is based on edge detection application property of Cellular Automata and I found a strong confirmation of all linear 2DCA rules showing this property although with a difference in quality. I have used a standard data matrix (Binary image) and applied one by one in all 512 linear rules of 2DCA with the help of MATLAB program. The standard data matrix is taken as an image of 128×128 pixels with multiple square and rectangular binary patterns as shown in Fig. (3). The following algorithm and flowchart were developed for handling the implementation of each rule on the above sample binary image matrix. The detailed results are presented in the next section.

The edge detection results studied here are compared with one another for their quality and limitations. The edge detection here was carried out with a similar process of 2DCA although with a difference of internal change of rule function. Out of 510 results, 280 produce full contours and other 198 rule functions hide horizontal or vertical edge information. There are 32 rule functions that produce corner detections and are highly efficient for the edge detection of circular and irregular image patterns.

4 Proposed Classification

On the application of rules numbered from 0 to 511, excluding Rule 0 (no combining interaction between the target cell and the neighbors) and Rule 1 (Interaction of the target cell with itself), all 510 rules show the property of edge detection, although of varied extent and quality. I have categorized these 510 rules into four basic classes on their behavioral difference in edge production on the boundaries of image features/elements. These four classes are:

drksujatha23@gmail.com

Table 1 Rule classes

S. no.	Class	Rules
1	Full	4, 5, 10, 11, 16–19, 28–31, 36–43, 50, 51, 60, 61, 64–79, 82–85, 90–93, 98–101, 106–109, 112–127, 130–133, 138–141, 144–175, 178–181, 186–189, 196–203, 210, 211, 220, 221, 228, 229, 234, 235, 240–243, 252–257, 262–265, 270–305, 310–313, 318, 319, 324–331, 336, 337, 350, 351, 358–361, 368–371, 380–383, 390–393, 400–403, 412–415, 420–427, 432, 433, 446–465, 470–473, 478–481, 486–489, 496–501, 504, 505, 510, 511 (280 Rules)
2	Horizontal	2, 3, 12, 13, 20–27, 32–35, 44–47, 52, 53, 58, 59, 192–195, 204–207, 212, 213, 218, 219, 226, 227, 236, 237, 244–251, 320–323, 332–335, 342–345, 352, 353, 366, 367, 372–379, 384, 385, 398, 399, 404–411, 417–419, 428–431, 438–441, 494, 495, 502, 503 (100 Rules)
3	Vertical	6–9, 48, 49, 62, 63, 80, 81, 86–89, 94–97, 102–105, 110, 111, 128, 129, 134–137, 142, 143, 176, 177, 182–185, 190, 191, 208, 209, 222, 223, 230–233, 258–261, 266–269, 306–309, 314–317, 338, 339, 348, 349, 356, 357, 362, 363, 388, 389, 394, 395, 434, 435, 444, 445, 466–469, 474–477, 482–485, 490–493, 508, 509 (98 Rules)
4	Corner	14, 15, 54–57, 214–217, 224, 225, 238–241, 346, 347, 354, 355, 364, 365, 386, 387, 396, 397, 436, 437, 442, 443, 506, 507 (32 Rules)

1. Full Detection Class
2. Horizontal Detection Class
2. Vertical Detection Class
4. Corner Detection Class

Out of these four classes, only one class produces edge detection on all sides of the image features although of varied quality. Table (1) presents the rule wise classification according to detection property of 2DCA rules in the above four categories. An observation of edge shift in all rules of each detection class was seen and recorded for which, correction needs to be applied in case of vital image data. Figure 4 demonstrates visually some rule detection results on binary data patterns to understand different types of detection classes.

In order to demonstrate the edge shift we applied the Rule 4(G1) on a low-resolution IR image of Dal Lake to obtain behavioral difference in edge production of its island (Rouplank) feature.

The results are presented in Table 2 with feature of island-encircled red. The edge shift of the feature from rectangular to parallelogram in enlarged form is presented in Fig. 5.

5 Conclusion

Comparing the results of various gradient-based Laplacian edge detectors with the above results, it has been observed that the 510 linear rules of 2DCA provide all operators, like that, used in Sobel, Robert, Prewitt or their improved form Canny.

drksujatha23@gmail.com

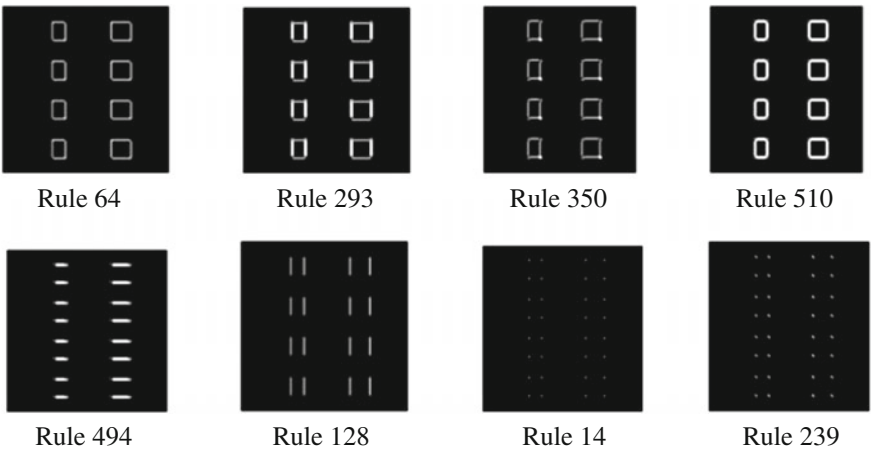


Fig. 4 Few 2DCA edge detection results

Table 2 Real image edge detection results

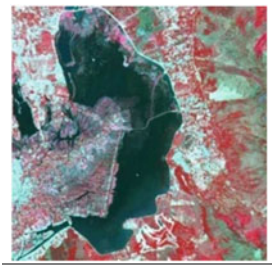
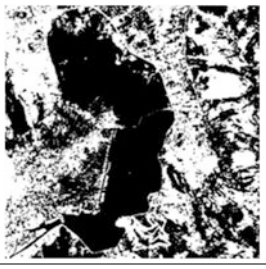
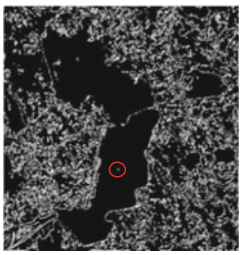
IR image	Binary image	Feature extraction using rule 4
		

Fig. 5 Red encircled feature enlarged



drksujatha23@gmail.com

Canny, that is termed as the optimal edge detector, type results are produced with 280 rule functions. Partial edge detection like that of Sobel, Robert, or Prewitt is produced by other 230 rule functions. Like that of finding edges in the first derivative in Laplacian procedures, 2DCA linear rule functions also produce result in the first iteration of the rule function execution. The important finding regarding the edge detection property of 2DCA rules is confirmed in this study. The experimentation reveals that the edge detection is produced in the first applied iteration of a particular 2DCA rule and further iterations lead to pattern evolution. The results achieved here have been presented in terms of experimental outcome and have also potentially solved the selection criteria problem of 2DCA rules for contour determination of digital image data. The edge shift in the contours of the images identified here needs to be taken as a separate assignment, in future, for finding correction procedures.

References

1. Neumann JV (1966) Theory of Self-reproducing automata. Scientific Research, Illinois: University of Illinois Press
2. Qadir F, Shah J, Peer MA, Khan KA (2012) Replacement of graphic translations with two-dimensional cellular automata, twenty five neighborhood model. IJCEM
3. Wani SJ, Qadir F, Peer ZA, Peer MA (2013) Quantum-dot cellular automata: theory and application. IEEE Explore. https://doi.org/10.1109/icmira.2013.113
4. Wani SJ, Qadir F (2017) 2D-cellular automata: evolution and boundary defects. IJSR, 1566–1570
5. Wolfram S (1984) Universality and complexity in cellular automata. Physca D 10:1–35
6. Wolfram, NH, Packard S (1983) Cellular automata complexity. J Stat Phys 38:901–946
7. Packard NH, Wolfram S (1986) Two dimensional cellular automata. J Stat Phys 38
8. Li W, Packard NH, Langton C (1990) Transition phenomena in cellular automata rule space. Physica D
9. Ohi F (2007) Chaotic properties of the elementary cellular automata rule 40 in Wolfram's Class I. Complex Syst 17:295–308
10. Michail M, Claude L, Jean CH (1997) Complexity classes in the two-dimensional life cellular automata subspace. Complex Syst 11:419–436
11. Choudhary, PP (2010) Classification of cellular automata rules based on their properties. IJCC, 8

drksujatha23@gmail.com

Augmented Intelligence: A Way for Helping Universities to Make Smarter Decisions



Manu Sharma

Abstract In the present specialized world, the word AI is itself is a significant word by its functionality, work efficiency, etc., it has totally changed the human's lifestyle, work processing, thinking ability, concurrency, behavior change, etc. Artificial Intelligence is better applied to helping individuals from propensities. It is a clinically effective and cost-effective tool via computerizing updates for practices. Lastly, it comes about for the future as a prompt future holds tremendous promise for hybrid systems. There is no formula for intuition means not every aspect can be replicated by an algorithm but it introduced "Augmented Intelligence," which best combines human and artificial intelligence to change human behavior. Popular visions of artificial intelligence often focus on robots and the dystopian future they will create for humanity, but to understand the true impact of AI, its skeptics and detractors should look at the future of cybersecurity.

Keywords AI (Artificial Intelligence) · Augmented intelligence
Human intuition and knowledge · Significance

1 Introduction

Artificial Intelligence plays an important role in computer science where a computer almost works like a human by better speed, more intelligence, and better-thinking process and with a good efficiency of working process as compared with the human's working speed and ability, and by this way, people are interacting with computers like they interact with the humans and they do not even face the problems in between to complete their task by using AI Technology. It is also assumed that there are some limitations on human thinking process and working but AI Technology machines do not have such types of limitation over thinking and

M. Sharma (✉)
Electronics & Communication Department, Gyan Vihar School of Engineering & Technology, Jaipur, India
e-mail: manu.sharma1988@yahoo.com; sarveshsingh@jvwu.ac.in

© Springer Nature Singapore Pte Ltd. 2019

V. S. Rathore et al. (eds.), *Emerging Trends in Expert Applications and Security*,
Advances in Intelligent Systems and Computing 841,
https://doi.org/10.1007/978-981-13-2285-3_11

drksujatha23@gmail.com

working. By day-by-day enhancements and working done on this technology to introduce some new techniques to do the similar task in update manner. So well and good to do the job in more efficient with speed and to maintain the work results on the scale which not only do these works but also improves the consistence of results, AI introduces Augmented Intelligence—which is a Artificial Intelligence presents and featured as a next-generation AI. In prior, it was not particularly utilized but rather, nowadays, individuals are utilizing this innovation for instruction, for clinics, and in different fields. This paper manages the importance of incorporating the advances together.

2 What Exactly the Augmented Intelligence Is

2.1 Augmented Intelligence

IBM has been exploring, creating, and putting resources into AI innovation for over 50 years. General society ended up noticeably mindful of a noteworthy progress in 2011 when IBM Watson won the memorable Jeopardy! Presentation on prime time TV. Since that time, the organization has progressed and scaled the Watson stage, and connected it to different businesses, including social insurance, back, trade, instruction, security, and the Internet of Things. At IBM, it is guided by the expression "augmented intelligence" as opposed to "artificial intelligence". It is the basic distinction between systems that upgrade and scale human ability as opposed to those that endeavor to recreate all of human knowledge and center on building handy AI applications that help individuals with all around characterized assignments, and all the while, uncover a scope of summed up AI benefits on a stage to help an extensive variety of new applications.

This new technology arises from AI itself: Augmented Intelligence means, in general, the use of technology to expand human information processing capabilities as it makes the better results that would be done and helps to solve humanity's grandest challenges. It takes a best of human intuition and imagination and combines with AI ability to maintain the scale, and access an early warning system for organization leading predict the things which might get wrong. Sometimes, AI as Augmented intelligence—moving from standardized to more complex environments requires AI that is more versatile (Fig. 1).

This new invention of AI has no motive to replace the human but to elaborate their thinking, whether in medicines, financial services or in any regulated business, etc. This emerging technology made the job easier by keep on depending on new research over continuous changing the regulations such as farmers with AI, surgeons with AI, and an art of using AI for complex Business Decisions.

drksujatha23@gmail.com

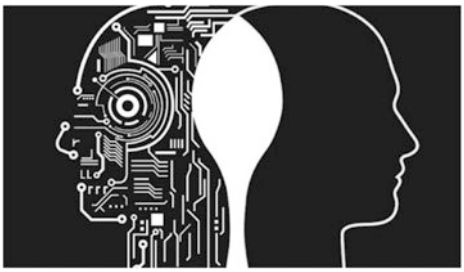


Fig. 1 Augmented intelligence in AI

2.2 Examples of Augmented Intelligence

The following are the examples of augmented intelligence which results successfully when implemented:

- Aircraft autopilot
- Legal discovery
- Healthcare
  - IBM Watson and University of North Carolina Cancer Center Study.
  - 1000 cancer patients.

    99% match on recommended treatment
    30% had valid alternative treatment missed by physicians (current clinical trial, recently approved, etc.)
- Computer Chess
  - The gap is narrowing.
  - But it is taking a long time.

2.3 Application Areas of Augmented Intelligence

1. **Augmented Intelligence in Business.**
2. **Manufacturing.**
3. **Transportation and Logistics.**
4. **Healthcare.**
5. **Agriculture.**

1. Augmented Intelligence in Business:

This innovation assumes an imperative part in business territory. The potential outcomes for business utilizes grow exponentially with just an impermanent problematic impact on the workforce instead of a lasting, heartbreaking impact. Rather than contemplating how autonomous robots can help your business, it

drksujatha23@gmail.com



Fig. 2 Augmented intelligence in business

moves toward how AI-driven advancements and devices can augment crafted by your workers to expand effectiveness and enhance profitability. Augmented intelligent systems have just been actualized at countless huge and independent ventures. By including AI-driven marketing automation technologies to upgraded information scientific systems, it is presumable business as of now utilizing some types of increased insight innovation (Fig. 2).

And this association between AI advances and individuals will even now disturb a few occupations; it will present opportunities for work development, higher profitability, and higher income for people and organizations in each industry. It will likewise open altogether new divisions of occupations for that work in increased insight frameworks.

The important terms using this advanced technology are as follows:

- This brings on new staff to the work.
- To avoid decision-making bottlenecks for the officers or workers in the organization.
- To manage customer relationships.
- Keep an eye on or to track all the financial conditions.

2. Manufacturing:

In manufacturing scale, this technology works as cooperative, intelligent robots that can securely work close by people and can deal with undertakings that are hard, dangerous, or monotonous will them help expand efficiency (Fig. 3).

3. Transportation and Logistics:

While there is a savage progressing open race among different innovation organizations and OEMs to drive towards completely self-sufficient vehicles, there is critical closer term opportunity from diminishing driver work stack in everyday driving circumstances, for example, roadways, decreasing mistake rates and mishaps in human-driven vehicles, and enhancing activity stream and fuel proficiency (Fig. 4).

drksujatha23@gmail.com

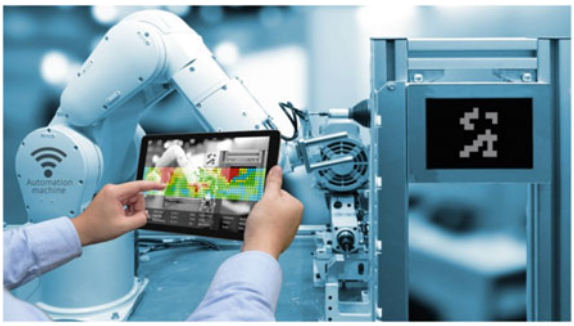


Fig. 3 Manufacturing of the AI machines



Fig. 4 Augmented intelligence in transportation and logistics

4. Healthcare:

As detailed earlier, it is known that all about the healthcare importance by this latest technology IBM Watson and University of North Carolina Cancer Center Study. 1000 cancer patients are relieved by this AI technique. 99% match on recommended treatment. 30% had valid alternative treatment missed by physicians (current clinical trial, recently approved, etc.) (Fig. 5).

5. Agriculture:

An assortment of cultivating robots, trim improvement methods, computerized water system systems, and nuisance cautioning systems will help increment agricultural productivity at high rates (Fig. 6).

Fig. 5 Augmented intelligence in healthcare



drksujatha23@gmail.com

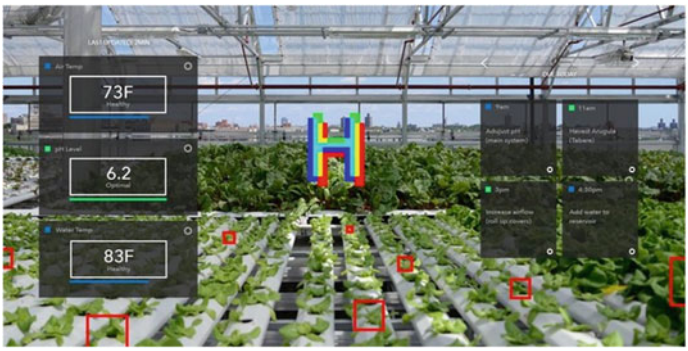


Fig. 6 Farming intelligence

3 Proposal for the Future Concern

Augmented intelligence systems are probably going to be imperative for years to come. What is more, this new innovation is actualized in each field effectively and performed best exertion and capacity with extraordinary outcomes, for example, in farming, the medical science, pharmaceuticals, associations, and so forth if this innovation works quick in these fields, it will without a doubt work in education field too. For instance, if in an association or faculty is not accessible when students need the questions to be tackled or any study-related queries to be solved, this enlarged insight will enable them to out. By bolstering the information of the specific subjects from essential level to larger amount. As Augmented intelligence works over human creative energy and thinking process, it will likewise takes a shot at student's state of mind and everyone of the inquiries asked by them that time. Many times students are not able to express or one can say by describing the query they want to solve at any place whether in front of a mentor or on the internet, Augmented intelligence must help them out with this problem also by understanding a student's state of mind.

This results in student's mind and knowledge improvement and may be the way of their thinking must be higher than before after implementation of this technology.

4 Conclusion

The future extension is completely clear, as individuals will rely upon AI with Augmented Intelligence, and these web search tools will soon vanish from everybody's life for any issues. Many research challenges are going ahead by utilizing discourse in Augmented Intelligence systems, with a few downsides like camera picture quality which are sufficiently bad that they can take great nature of pictures

drksujatha23@gmail.com

from a distance and additionally in dark room. Subsequent to explaining these comparable disservices, it can be taken and after that, better outcomes can be derived.

drksujatha23@gmail.com

A Multiple String and Pattern Matching Algorithm Using Context-Free Grammar



Sarvesh Kumar, Sonali Singh, Arfiha Khatoon and Swati Agarwal

Abstract Various Substring Pattern Matching Algorithm introduced here is executed in two stages. The main stage is preprocessing in which a n-ary tree as structure is built for the given content information and GNF (i.e., Griebach Normal Form) is made for given Context-Free Grammar. The second stage called seek stage takes as info a n-ary tree structure and Griebach Normal Form of given Context-free linguistic use is built in stage 1, and yields those strings that match both the content information and the setting free language structure. The calculation proposed here has the favorable position that it can recover any number of examples in the meantime. This finds wide applications in bioinformatics, data recovery, and prerequisites particular phase of programming life cycle advancement.

1 Introduction

The calculation proposed here is extremely useful in recovering the substrings of given setting free dialect in a simple and productive way. The claim to fame of the calculation proposed here is that it does not require any pushdown automata for the acknowledgment of the strings produced by given setting free linguistic use. The time taken to recover the substrings is same for same size of content database and same setting free sentence structure that implies look time free of the substance of content database Planguage. The time taken by our calculation is free of the measure of examples, which overcomes the real restriction which is available in the

S. Kumar (✉) · S. Singh · A. Khatoon · S. Agarwal
Jayoti Vidyapeeth Women's University, Jaipur, Rajasthan, India
e-mail: sarveshsingh@jvwu.ac.in

© Springer Nature Singapore Pte Ltd. 2019

V. S. Rathore et al. (eds.), *Emerging Trends in Expert Applications and Security*,
Advances in Intelligent Systems and Computing 841,
https://doi.org/10.1007/978-981-13-2285-3_12

drksujatha23@gmail.com

past calculations. The language structure guidelines of formal linguistic uses are utilized to create designs. Common English Language can likewise be gotten from a setting free punctuation with fitting preparations. For a given language structure G, the quantity of strings created by that language structure is vast that implies |L (G)| is ∞. Parsing is a crucial idea identified with the syntactic approach whose goal is to decide whether the info design is linguistically all around shaped in the setting of the given language structure. Parsing is for the most part achieved by parsers. In the inquiry calculation displayed, here, we discover the crossing point of the strings shown in the content database and the dialect of the given sentence structure, for that we have not utilized any parser. Numerous substrings can be gotten at the same time utilizing the hunt calculation introduced here.

2 Processing Phase

2.1 Preprocessing of Text Data

The content information is preprocessed with the goal that hunt time is enhanced. Content information is only the arrangement of strings. The content information is spoken to with a structure like n-ary tree. The content information is taken, each string is passed to tree and for each string, the tree is changed. The tree structure is like n-ary tree where n speaks to the number of unmistakable images in the given content information, in the introduced work. The hub structure for tree is given by Madhuri et al. [1].

Node Structure of every hub of the tree has the accompanying fields.

(i) Information Field: This field contains data.
(ii) Pointers to the kid hubs: The pointers that hold the addresses of tyke hubs (the No. of kids is at max is n).
(iii) Flag: Banner is utilized to perceive whether the specific string acquired by connecting the strings from the root up to the current hub in the left to right grouping is available as a component in the database or not. The banner of a hub is 1 if and just if the hang to the comparing hub is contained in the database. The calculation for the development of database tree given by

Preprocessing of the setting free syntaxes:
The question is given as setting free syntaxes. The given setting free linguistic use is changed over into a Griebach Normal Form which is likewise a unique situation free syntax. The dialect created by both the setting free syntax and its Griebach typical frame is same. The preprocessing of the setting free linguistic uses

drksujatha23@gmail.com

expands the productivity of looking. The transformation happens in five stages. They are:

1: Reduction of useless images.
2: Reduction of null preparations.
3: Reduction of unit productions.
4: Transformation to Chomsky like typical frame.
5: Transformation to Griebach typical shape.

Pursuit Phase

A context-free grammar structure G comprises of the following four substances:

1. The arrangement of terminals or primitive images is signified by T. In numerous applications, the decision of the terminal set is troublesome and has a substantial part of workmanship rather than science. T is limited.
2. The arrangement of nonterminal images or factors which are utilized as transitional amounts in the age of result comprising exclusively of terminal images. This set is meant as V and it is additionally limited.
3. The arrangement of creations or generation decides that permit the past substitutions. It is this arrangement of creations combined with terminal images that chiefly gives the language structure its structure. The arrangement of preparations is meant by P.
4. The beginning or root image indicated by S has a place with V.

The punctuation G is meant formally as G = (T, V, P, S). A dialect L is said to be a Context-Free Dialect (CFL), if its linguistic use is Context-Free. The creation rules P, is utilized to produce sentences that comprise of direct or 1D series of terminals. The length or, on the other hand, the number of images in string s is signified by |s|. The purge string is meant by ε. The extent of purge string is 0. Convergence of the terminals and the factors is void set ($V \cap T = \Phi$).

3 Algorithm for Looking Through the Database Tree Utilizing Context-Free Grammars

Input:

1. n-ary tree portrayal of content information.
2. Griebach Normal Form of given context-free grammar structure.

Yield: List of information base tree pointers that matches the given context-free grammar structure struct lst *explore_cfg[char *prod,structlst *db_lst]

drksujatha23@gmail.com

```
//exploring CFG that  is in GNF
{
//'prod' is a string that is as terminal took after by non terminals
//'db_list' is the rundown of database pointers in the
arrangement of single linkedlist
if(is terminal(prod[0]))
{
{
coordinate the terminal in 'prod[0]' to the present
hub in the 'db_list'
    if(there is no result)
    {
    Return[null]
    }
    Else
    {
    the following hub to be looked in the database
    if(the hub in the 'db_list' to be looked is NULL)
{

if ((production+1) = = “NULL”)
{
return (db_list);
}
    else
    {
    Return(NULL);
}
}
else
    {
Temp_list=search_cfg(production+1,db _list)
return(Temp_list);
}
}
}
else
{
Temp_list1=NULL;
for each correct hand side creation
'temp_prod' of non-terminal 'prod[0]'
{
Annex the returned list from
'search_cfg(temp_prod,db_list)' to
Temp_list1
    }
Temp_list3=NULL;
for each 'Temp_list2' in 'Temp_list1' connected list
{
Add the returned list from
'search_cfg(prod+1,Temp_list2)' to
Temp_list3
```

drksujatha23@gmail.com

```
            }
            return(temp_list3);
            }
        }
        void CFG_substring_match()
        {
        db_list db_ls_2, db_ls_3;
        for every hub 'node_1' in data base base tree
        {
        for each character position 'pos_1' in 'node1'

    {

    db_list db_ls_1;

    db_ls_1 = make new db_list;

    db_ls_1->pos = pos_1;

    db_ls_1->data = node_1;

    db_ls_3 =

    search_cfg(db_ls_1,Starting_charact

    er_of_CFG);

    Add 'db_ls_3' to 'db_ls_2';

    return(db¬_ls_2);

    }

    }

    }
```

This is the calculation that discovers every one of the substrings for the given database that fulfills the context-free grammar. The routine “search_cfg” for finding the complete strings is utilized as a part of this calculation inside. For every last hub in the database, we will call the “search_cfg” schedule with the goal that every one of the substrings likewise recovered.

drksujatha23@gmail.com

4 Applications

Necessities are the premise of the frameworks designing life cycle activities, however, making a decent set of necessities is truly troublesome errand. A few troubles can be diminished through the use of a setting free sentence structure for necessities to lessen the many-sided quality of necessities elicitation. Building up the linguistic use included a merging of software engineering and common dialect that yielded helpful bits of knowledge into the idea of necessities. The linguistic use was produced to enable a case-based appraisal framework for necessities. Bioinformatics includes the utilization of strategies counting connected arithmetic, informatics, measurements, software engineering, counterfeit consciousness, science, and organic chemistry to take care of natural issues as a rule at an atomic level. The calculation introduced here can be used to recognize particular examples of amino acids in the DNA grouping and in this way, it can help in the diagnosis of certain maladies. For this situation, the terminal set comprises of {a, c, g, t}.

5 Conclusion

The calculation proposed here is extremely useful in recovering the substrings of given setting free dialect in a simple and productive way. The claim to fame of the calculation proposed here is that it does not require any pushdown automata for the acknowledgment of the strings produced by given setting free linguistic use. The time taken to recover the substrings is same for same size of content database and same setting free sentence structure that implies look time free of the substance of content database.

References

1. Smale S (1985) On the efficiency of algorithm analysis. Bull Am Math Soc 13:2
2. Abarbanel RM, Brutlag DL (1984) Rapid searches for complex patterns in biological molecules. Nucleic Acids Res 12(1):263–280
3. Hanna FH (1976) Automata theory: an engineering approach. Crane, Russak & Company Inc
4. (2007) A fast multiple pattern matching algo using cfg & tree model. IJCSN 7(9)
5. Smale S (1990) Some remarks on the foundation of numerical analysis. 32(2)
6. Luger G, Stubblefield WA (1998) Artificial intelligence, structures and strategies for complex problem solving
7. Martin J (2000) Requirements mythology: shattering myths about requirements and the management thereof. In: Proceedings of the 2000 INCOSE symposium

drksujatha23@gmail.com

A Review of Machine Translation Systems for Indian Languages and Their Approaches



Dipal Padhya and Jikitsha Sheth

Abstract Translation is the obvious requirement to abolish the communication barrier. The barrier may occur while knowing different languages and prevent from sharing the information. This paper provides a survey of Machine Translation (MT) systems developed for Indian languages. It also provides an idea regarding the approaches and evaluation techniques used for translation. From this paper, a researcher can have a glance regarding the work carried out for Indian languages and enhance the work from where it stops. Until now, a lot of work has been carried out for MT. Some systems are developed for general domain whereas others are for specific domain like administrative documents translation, news translation, children stories, weather narration and conference papers, etc., and still some languages require more attention.

1 Introduction

India is the home to several hundred languages and has 22 constitutionally recognized official languages [1]. It contains 122 major languages and 1599 other languages [1]. Hindi and English are the official languages used by the central government. State governments use their respective official languages for official work. Translation is required in order to pass the work of state government to the central government. That translation should be in official languages only.

Due to this diversity in languages, it is almost impossible for everyone to know all the languages. Even the newspapers are published in various languages. So to overcome all these limitations, translation is required.

MT is a subfield of computational linguistics that translates the text or speech from one language to another [2]. The MT field appears in Warren Weaver's

D. Padhya (✉) · J. Sheth
Uka Tarsadia University, Bardoli, Gujarat, India
e-mail: dipaldpadhya@gmail.com

J. Sheth
e-mail: jikitsha.sheth@utu.ac.in

© Springer Nature Singapore Pte Ltd. 2019

V. S. Rathore et al. (eds.), *Emerging Trends in Expert Applications and Security*,
Advances in Intelligent Systems and Computing 841,
https://doi.org/10.1007/978-981-13-2285-3_13

drksujatha23@gmail.com

Memorandum on translation in 1949 [2]. Yehoshua Bar-Hillel, the first researcher in the field, begins his research in 1951 [2]. In India, the efforts for MT starts from the mid-80s and early 90 s and are still going on.

This paper focuses on various MT systems developed for Indian languages. To understand those MT systems, the knowledge of approaches is a must. In the next section, various MT approaches are discussed.

2 Machine Translation Approaches

The following subsections provide the description of MT approaches with utilized MT systems.

2.1 Rule-Based Machine Translation (RBMT)

It parses the source text and produces an intermediate representation, which may be parse tree or some abstract representation [3]. English to Sanskrit [4], English to Urdu [5], EtranS [6], English to Marathi [7], English to Marathi [8], English to Marathi [9], English to Kannada [10], and TranSish [11] MT systems utilized this approach for development.

Direct Machine Translation (DMT)

As the name suggests, direct MT system provides direct translation, i.e., no intermediate representation is used in [3]. Punjabi to Hindi [12], Hindi to Punjabi [13], Hindi to Punjabi [14], English to Devanagari [15], and English to Sanskrit [16] MT systems used this approach for development.

Transfer-Based Machine Translation (TBMT)

A transfer-based MT system involves three stages. The first stage makes analysis of the source text and converts it into abstract representations, the second stage converts those into equivalent target language, and the third generates the final target text [3]. Telugu to Tamil [17], Bengali to Hindi [18], Punjabi to English [19] and Malayalam to English [20] MT systems used TBMT for development.

Interlingual Machine Translation (IMT)

The IMT operates over two phases: analyzing the source language text into an abstract universal language-independent representation of meaning, i.e., the

drksujatha23@gmail.com

Interlingua, which is the phase of analysis; generating this meaning using the lexical units and the syntactic constructions of the target language [3]. English to Sanskrit [21] and English to Bengali [22] used this approach for implementation.

2.2 Statistical Machine Translation (SMT)

In SMT, translations are generated based on statistical models [3]. English to Malayalam [23], English to Urdu [24], English to Kannada/Telugu [25], English to Sanskrit [26], Punjabi to English [27] and English to Urdu [28] utilized SMT for development.

2.3 Example-Based Machine Translation (EBMT)

EBMT is characterized by its use of bilingual corpus with parallel texts as its main knowledge, in which translation by analogy is the main idea. There are two modules in EBMT: example retrieval and adaption [3]. Malayalam to English [29] and English to Hindi [30] used EBMT for development.

2.4 Hybrid Machine Translation (HMT)

HMT is characterized by the usage of multiple MT approaches within a single MT system. AnglaHindi (English to Hindi) [31], ANUBAAD (English to Bengali) [32], Bengali to Hindi [33], English to Punjabi [34], Urdu to English [35], English to Malayalam [36] and English to Sanskrit [37] used this approach for the development.

From the above discussion, the brief knowledge of Indian languages can be gained for which MT work has been carried out. In the next section, those MT systems are discussed in detail.

3 Literature Survey

The Indian MT systems are discussed below based on their utilized approaches.

3.1 MT Systems Based on RBMT

In English-Sanskrit MT system, Artificial Neural Network (ANN) was used with rule-based model and highest achieved score is “1.00 by METEOR” [4]. Same way ANN was used with RBMT in English-Urdu MT system and achieved score is

drksujatha23@gmail.com

"0.86 by f-score" [5]. The translation was based on the formulation of Synchronous Context Free Grammar (SCFG) in Etrans and highest achieved accuracy is 99% [6]. Artificial Intelligence can also be added with dictionary for increasing the accuracy and performance. It was implemented in Sanskrit to English [11]. The [9] MT system was built for English-Marathi language. It provided word sense disambiguation model and used stanford parser. The [7] MT system was also built for English-Marathi. But it was developed for improving the performance and provide spell checking, translation of idioms and sentiment analysis as newly added functionality. Morphological generator was used to handle complex morphology in English-Kannada MT system [10]. English to Marathi [8] was developed to handle assertive sentence only.

MT Systems Based on DMT

Other than dictionary, the additional Vichheda Module was added in English to Sanskrit [16] MT system. This module is responsible for identifying and forming words using word generator. Other than sentence translation, an additional Email translation functionality is available in Punjabi to Hindi [12] MT system and provides 90.6% accuracy. The same functionality is available for Hindi to Punjabi translation [13]. But the difference is here, input is converted into unicode character first and then it is translated with 95% accuracy, which is more compared to [12]. An advance version of Hindi to Punjabi is also created by the same authors [14]. However, here the system is not web-based tool. The English to Devanagari MT system is the example of transliteration where the labels/words/phrases like "Name of the father" is converted to English-Marathi, English-Hindi and English-Gujarati [15].

MT Systems Based on TBMT

Other than just providing translation, the Punjabi to English MT system provides synthetization and transliteration using bilingual dictionary [19]. Same way in Telugu-Tamil [17] MT system, cross-linguistic variations are handled. The Bengali to Hindi [18] MT system converts the sentence with the help of lattice-based data structure. The Malayalam to English [20] MT system works for sentences, which contain up to two adverbial or adjectival clauses, which is commonly found in Malayalam texts.

MT Systems Based on IMT

Rather than giving focus on generating Interlingua, the English to Sanskrit [21] MT system focused on lexical parser and semantic mapper. Approximate Lexical Meaning Mapping (ALMM) can be used with CFG. The English to Bengali MT system is the example of it [22].

drksujatha23@gmail.com

3.2 MT Systems Based on SMT

The Punjabi to English [27] MT system provides transliteration along with translation with the help of unigram, bigram, trigram, four-gram, five-gram, and six-gram and achieved accuracy is 97%. The English to Kannada/Telugu [25] MT system also did transliteration, but with the help of bilingual corpus, transliteration model, etc. and achieved accuracy is 96.6%. Hence, it can be stated that the difference regarding accuracy is very low. Moses is an SMT tool that allows to train translation models automatically for any language pair. The English to Urdu [28] MT system uses Moses translation setup with language modeling toolkit IRSTLM and achieved BLEU score is 37.10. In English to Urdu [24] MT system, Moses toolkit with Giza++ was used. The alignment model with category tags was used in English to Malayalam [23]. Statistical machine decoder was used in English to Sanskrit [26].

3.3 MT Systems Based on EBMT

Here, the examples containing source and target languages are stored. In Malayalam to English [29] MT system, the comparison of source and target strings was carried out with the help of inbuilt functions of MATLAB and achieved accuracy is 75%. In the same way, similarity matrix, training matrix, and tagging matrix were used for comparing English to Hindi [30] MT system and achieved accuracy is 86%.

3.4 MT Systems Based on HMT

In English to Sanskrit [37] MT system, translator and synthesizer were created by combining RBMT with EBMT. Whereas in Urdu to English [35] MT system, the comparison between RBMT, EBMT, and SMT was shown. The English to Punjabi [34] MT system first used DMT for translation. When the translation is not achieved with it, then only it will be performed by EBMT. A Translation Memory (TM) is a linguistic database that continually captures translations for future use. It can be used with SMT to save the time and enlarge the performance of the system. The English to Malayalam [36] MT system first searched sentence/word/phrase into TM and if not available in it, then only it does translation using SMT. AnglaHindi [31] was designed to cater compound, complex, imperative, interrogative, and other constructs such as headings, etc., using RBMT and EBMT. ANUBAAD [32] works with monosemous words and can translate simple and compound sentences using TBMT and EBMT.

drksujatha23@gmail.com

As per the study carried out, it has been found that 47.5% work of MT was conducted using RBMT while 17.5% work was conducted using SMT and 5% work used EBMT approach for development. The remaining work was carried out by HMT approach. Therefore, it can be stated that "The usage of RBMT was more compared to other approaches". Various tools/technologies were employed to build systems but most employed technology is Java. Human or automated system like BLEU, METEOR, NIST, etc. can be used for evaluating the MT system.

4 Findings

The error rate of words in Punjabi to Hindi [12] MT system is comparatively lower than that found for other languages.

POS (Part Of Speech) information can be added into a bilingual corpus while doing alignments. It removes insignificant alignments and adds the capability to the system to translate unseen sentences. As per the study conducted, the systems utilizing statistical model without POS tagged information are not able to translate unseen sentences. Thus, it can be stated that to accumulate the efficiency of the MT system using statistical model, the corpus with POS tagged information should be used.

Word Sense Disambiguation is the common problem almost for all the natural languages. While doing translation, it should be handled. For handling it, n-gram can be used. Until now most utilized n-grams are of size 2(bigram) and of size 3 (trigram). They help in resolving ambiguity by searching the words into n number of sequence. Besides handling it, new words must be added into database those are coming during translation. Those words must be added to the lexicon only after correcting it if, not properly transliterated.

The possibilities for work extension is almost there for all the language pairs. But the easiness and possibility may be affected by several parameters. For example, there are some systems developed for Sanskrit to English language. They mainly focused on parasmaipada and Lat lakāra (present tense). These systems can be easily extended for atmanepad, past tense, and future tense for Sanskrit to English. But the same task will become challenging when the language pair is Sanskrit to Gujarati due to sentence order variation. Likewise while working with speech recognition, it will again become a challenging task when the language pairs are different in sound.

Every MT system must be evaluated for checking its accuracy and improving the system's performance. Almost all the MT systems developed for Indian languages provide good accuracy but the accuracy for Sanskrit systems is comparatively higher than that found for other languages.

There is a big scope for Sanskrit language, even being an ancient language and mother of all the languages very limited work has been carried out. Due to it is phonetically correct and rich in morphology, NASA also finds it as a most friendly language to work with computers.

drksujatha23@gmail.com

5 Conclusions

In this paper, research on various MT systems for Indian languages was conducted. Most of the systems are designed for specific domain while others are general systems. For the same ordered languages, it is best to use DMT. EBMT and SMT are best suited when the large corpus is available. Usage of RBMT is better when no corpus is available. But here, the knowledge of deep linguistic is must for constructing rules and providing good quality. Almost all the Indian languages are phonetically correct and have the scope for future enhancement.

References

1. https://en.wikipedia.org/wiki/Languages_of_India
2. https://en.wikipedia.org/wiki/Machine_translation
3. Siddiqui T, Tiwary US (2008) Natural language processing and Information retrieval, Ox ford University Press
4. Mishra V, Mishra RB (2010) ANN and rule based model for English to Sanskrit MT, Department of Computer Engineering, IT-BHU, Varanasi, U.P., India
5. Khan S, Mishra RB (2011) Translation rules and ANN based model for English to Urdu MT. INFOCOMP 10(3):36–47
6. Bahadur P, Chauhan DS, Jain AK (2012) EtranS–a complete framework for English to Sanskrit MT. IJACSA Special Issue, pp 52–59
7. Pishartoy D, Sidhaye P, Wandkar S (2012) Extending capabilities of English to Marathi machine translator. In J Comput Sci 9(3)
8. Abhay A, Anita G, Paurnima T, Prajakta G, Priyanka K (2013) Rule based English to Marathi translation of assertive sentence. IJSER 4(5):1754-1756. ISSN 2259-5518
9. Gajre GV, Kharate G, Kulkarni H (2014) Transmuter: an approach to rule-based English to Marathi MT. Int J Comput Appl 98(21)
10. Basavaraddi CCS, Shashirekha HL (2014) A typical MT system for English to Kannada. Int J Sci Eng Res 5(4)
11. Upadhyay P, Jaiswal U, Ashish K (2014) TranSish: translator from Sanskrit to English-a rule based MT. Int J Curr Eng Technol
12. Josan GS, Lehal GS (2008) A Punjabi to Hindi MT system. In: Coling 2008: companion volume posters and demonstrations, Manchester, pp 157–160
13. Goyal V, Lehal GS (2010) Web based Hindi to Punjabi MT system. J Emerg Technol Web Intell 2(2), May 2010
14. Goyal V, Lehal GS (2011) Hindi to Punjabi MT system. In: Proceedings of the ACL-HLT 2011 system demonstrations, pp 1–6, Portland, Oregon, USA, 21 June 2011
15. Dhore ML, Dixit SX (2011) English to Devnagari translation for UI labels of commercial web based interactive applications. Int J Comput Appl 35
16. Rathod SG, Sondur S (2012) English to Sanskrit translator and synthesizer. Int J Emerg Technol Adv Eng 2(12)
17. Parameswari K, Sreenivasulu NV, Uma Maheshwar Rao G, Christopher M (2012) Development of Telugu-Tamil bidirectional MT system: a special focus on case divergence. In: Proceedings of 11th international Tamil internet conference, pp 180–191
18. Chatterji S, Sonare P, Sarkar S, Basu A (2011) Lattice based lexical transfer in Bengali Hindi MT framework. In Proceedings of ICON-2011, India

drksujatha23@gmail.com

19. Batra KK, Lehal GS (2010) Rule based MT of noun phrases from Punjabi to English. Int J Comput Sci 7(5), September 2010
20. Nair L, Peter D, Ravindran R (2012) Design and development of a Malayalam to English translator-a transfer based approach. IJCL 3(1)
21. Barkade VM, Devale PR, Patil SH (2010) English to Sanskrit machine translator lexical parser and semantic mapper. NCICT-2010
22. Shibli A, Humayun K, Musfique A, Noman KM (2013) English to Bengali MT using context free grammars. Int J Comput Sci 10(3), No 2
23. Sebastian MP, Sheena Kurian K, Kumar SG (2009) English to Malayalam translation: a statistical approach
24. Ali A, Siddiq S, Malik M (2010) Development of parallel corpus and English to Urdu Statistical MT. Int J Eng Technol 10(5)
25. Reddy MV, Hanumanthappa M (2011) English to Kannada/Telugu name transliteration in Clir: a statistical approach, Bioinfo Publications
26. Warhade S, Devale P, Patil S (2012) English-to-Sanskrit statistical MT with ubiquitous application. Int J Comput Appl 51(1)
27. Kumar P, Kumar V (2013) Statistical MT based Punjabi to English transliteration system for proper nouns. IJAIEM 2(8), pp 318–320
28. Ali A, Hussain A, Malik MK (2013) Model for English-Urdu statistical MT. World Appl Sci J 24(10): 362–1367
29. Anju ES, Manoj Kumar KV (2014) Malayalam to English MT: an EBMT system. IOSR J Eng (IOSRJEN) 04(01):18–23
30. Sinhal RA, Gupta KO (2014) A pure EBMT approach for English to Hindi sentence translation system. I.J. Mod Edu Comput Sci 7:1–8
31. Sinha RMK, Jain A (2003) AnglaHindi: an English to Hindi machine-aided translation system. Indian Institute of Technology, Kanpur, India
32. Bandyopadhyay S et al (2004) ANUBAAD–a hybrid MT system from English to Bangla. In: Symposium on Indian Morphology, Phonology & Language Engineering, pp 91–92
33. Chatterji S, Roy D, Sarkar S, Basu A (2009) A hybrid approach for Bengali to Hindi MT. In: Proceedings of ICON-2009, pp 83–91
34. Kaur H, Laxmi V (2013) A web based English to Punjabi MT system for news headlines. Int J Adv Res Comput Sci Softw Eng 3(6):1092–1094
35. Habib A, Malik AA (2013) Urdu to English MT using bilingual evaluation understudy. Int J Comput Appl 82(7):5–12
36. Nithya B, Joseph S (2013) A hybrid approach to English to Malayalam MT. Int J Comput Appl 81(8):11–15
37. Rathod SG (2014) MT of natural language using different approaches: ETSTS (English to Sanskrit translator and synthesizer). Int J Comput Appl 102

drksujatha23@gmail.com

Energy-Efficient Cloud Computing for Smart Phones



Nancy Arya, Sunita Chaudhary and S. Taruna

Abstract In recent years, smart phones are becoming more popular day by day. According to the IDC study, the market of smart phones will reach 2.0 billion by 2019. Smart phones are now capable to support a high range of applications such as making voice and video calls, playing 3D games, etc., many of which computational intensive which results shortened battery life and poor performance. Though smart phones of current generation have powerful resources, energy efficiency is one of the main constraints for smart phones. Energy efficiency of mobile devices can augment by only 5% per annum by using the avant-garde technologies. Thus, it is a major challenge to improve energy efficiency and performance of smart phones. Mobile Cloud Computing (MCC) employs computational offloading to overcome the issues related to storage, energy, and performance of mobile devices. Thus, this paper emphasizes on computational offloading or augmented execution to augment the performance and energy efficiency of resource-demanding mobile applications in the cloud-assisted environment.

Keywords Mobile cloud computing · Energy efficiency · Augmented execution Computational offloading

N. Arya (✉) · S. Chaudhary
Department of Computer Science, Banasthali Vidyapith, Vanasthali, Rajasthan, India
e-mail: nanarya1@gmail.com

S. Chaudhary
e-mail: sunitaburdak@yahoo.co.in

S. Taruna
Department of Computer Science, JK Lakshmipat University, Jaipur, Rajasthan, India
e-mail: staruna71@yahoo.com

© Springer Nature Singapore Pte Ltd. 2019

V. S. Rathore et al. (eds.), *Emerging Trends in Expert Applications and Security*,
Advances in Intelligent Systems and Computing 841,
https://doi.org/10.1007/978-981-13-2285-3_14

drksujatha23@gmail.com

1 Motivation

Ubiquity and mobility are the two main features which facilitate with many network services. Ubiquities mobile network and cloud computing engender a new paradigm known as mobile cloud computing. It overcomes the obstacles of performance, security, and environment. Now, smart phones are used for various multimedia applications which drain the battery swiftly. Processing speed and storage capacity is inversely proportional to power saving which hinders the actual execution and performance of applications on device. These restrictions may be alleviated by offloading which send computational intensive data to cloud for execution and then receive the results back by cloud. In offloading, the applications that can adaptively be split and components offloaded are known as elastic mobile applications [1]. These applications decides at run time which parts of the application are executed on cloud. Several computational offloading techniques exist to augment performance and save battery of smart phones in cloud environment but mostly are inadequate for calculating the additional overhead of components migration at runtime. Thus, it is essential to focus on the processing of resource-demanding mobile applications so that performance and energy efficiency can be improved. The paper is organized into three sections excluding the motivation. Section 2 is related to work. Section 3 is communication strategy includes proposed solution and case study. Finally, conclusion is given in Sect. 4.

2 Related Work

Although, a lot of research work have been done in cloud computing to make computational offloading practically adopted. Energy is a fundamental factor for battery powered devices [2]. The µcloud model [3] is an energy-based offloading model for mobile applications in cloud assisted environment which decoupled the application components from each other. Niu, Song et al. in [4] proposed an algorithm named as Energy-Efficient Multisite Offloading Algorithm (EMSO) that works on workflow to check the tasks, size of tasks, and tasks' dependency of mobile application to improve the energy efficiency. Cuervo et al. [5] proposed MAUI as energy and method based offloading technique for mobile applications that support a semi-dynamic partitioning. Wang et al. [6] proposed an efficient heuristic algorithm, which minimizes the execution delay as well as energy consumption of joint optimization applications. Clone cloud model [2] generates a device clone in cloud environment. Data is transferred to the clone for execution to improve performance [7]. Yang in [8] has proposed an application partitioning framework to improve the throughput and speed to enhance the performance. Saad et al. [7], proposed an application development model to improve energy and

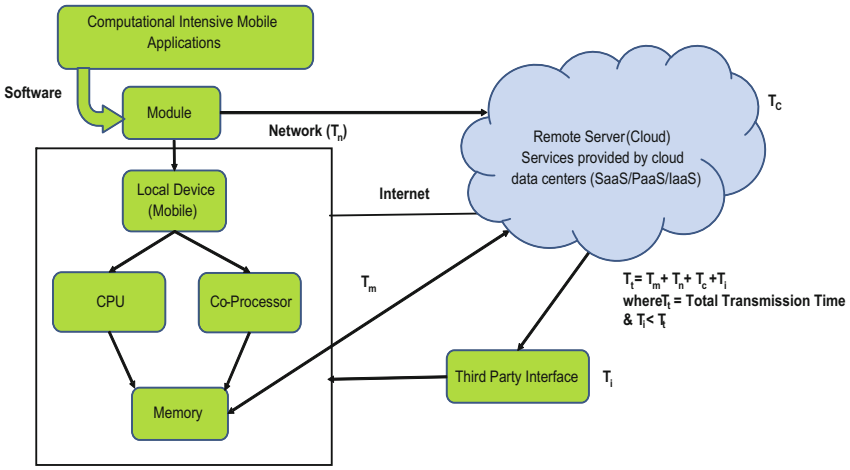
drksujatha23@gmail.com

performance both for matrix multiplication application by offloading on cloud. Shiraz et al. [9] proposed an energy-efficient model for the execution of the intensive mobile application in which minimal instances of the computational intensive application is offloaded at runtime [9].

3 Communication Strategy

Conventional computational offloading techniques apply adaptive algorithms to dynamically distribute intensive parts to the cloud which increases the power usage and turnaround time. There is a need for a framework to minimize the response time and battery consumption as well. The problem aims to utilize application processing services of cloud with the migration of least instances of the computational intensive application at runtime. Minimizing the overhead at runtime results in the minimum cost of power usage.

Proposed Solution:
The instances of the computational intensive application can be minimized at runtime by employing computational task offloading as a primary offloading instead of migration of intensive component. Thus, the cost of energy consumption will reduce. To achieve the goal, three points have to consider what instances of intensive application to offload, how to offload the components and finally, where to offload the components? To improve the performance and energy efficiency, least intensive part of the mobile application executes locally and most intensive instances of the application that takes more energy and high execution time will migrate to the cloud node. This partitioning decision depends on various input



1 Flow of application execution

drksujatha23@gmail.com

conditions of cloud node and mobile device such as the type of application, execution time on local device, resource utilization, CPU load, bandwidth, battery on device, and type of connection, etc. The flow of application execution is shown in Fig. 1.

The application data which is stored in memory is either executed on local device by CPU or on resourceful servers as shown in the figure. If the application data needs more energy and time on the device, data is then offloaded to the cloud servers. Execution time and resources utilization for the mobile device and turnaround time of the application on the cloud servers are analyzed. Sending the most intensive instances of the application for remote processing at runtime will reduce the increased overhead that results augmented energy efficiency and performance. Time needed is indirectly proportional to the performance and energy efficiency. Thus, total time (T_t) taken by data to migrate from mobile to cloud can be calculated by following formula:

$$\begin{aligned}\textbf{Total Time } (\mathbf{T_t}) = {} & \textbf{time spent on memory to fetch data and to writing back } (\mathbf{T_m}) \\ & + \textbf{transmission time on network } (\mathbf{T_n}) + \textbf{total timespent on cloud } (\mathbf{T_c}) \\ & + \textbf{time taken on third party interface } (\mathbf{T_i})\end{aligned}$$

Third party interface is any third party vendor that provides the communication between mobile and cloud. Time of interface should not be greater than total transmission time ($\mathbf{T_{i\,<}\,T_{t)}}$. By reducing this total amount of time, performance and energy efficiency may be increased gradually.

Case Study:
The proposed model evaluated with the matrix multiplication application on the platform of android Xiaomi Redmi note 4 phone system in the simulated environment to check the energy usage on various stable and unstable Internet connections of various bandwidths. Redmi note 4 is equipped with an Octa-core Qualcomm snapdragon 625 Max 2.0 Ghz processor, and 3.00 GB RAM. It supports Wi-Fi IEEE 802.11 a/b/g/n interface. The Redmi note 4 runs Android 6.0 marshmallow and is powered by 4100 mAh nonremovable battery with standby time of up to 775 h on 3G. Power Tutor Tool is utilized to measure the power consumption in application processing and Dalvik Debug Monitor System (DDMS) is used as a monitoring tool for the measurement of resources utilization. By taking multiple readings with each connection number of times, lot of variation generates in the

1 Test cases summary

Connection	Time elapsed (ms)	Consumed energy (%)
4G	37520	6
3G	36789	6
Wi-fi	5096	1
Edge	67596	6
Device	1295	2.5

drksujatha23@gmail.com

execution time as edge; 3G and 4G connections are not stable with respect to place. In contrast, there were negligible variations on stable connections in terms of speed and bandwidth such as Wi-Fi and on device (Table 1).

4 Conclusion

Mobile cloud computing helps to save power and execution time by migrating the mobile applications from mobile to cloud. This paper concludes that the proposed solution emphasized for the execution of computational intensive mobile applications in the cloud assisted environment. The solution utilized the cloud services for processing of applications with migration of least instances at runtime to reduce the increased overhead and additional energy consumption. Energy efficiency and performance can be augmented by reducing an increased overhead. Total required time is indirectly proportional to the performance and energy efficiency. By reducing the total amount of transmission time, performance and energy efficiency may be increased gradually. Offloading the tasks on the cloud by using the stable high speed Internet connection is also a good solution to save the energy and time of mobile applications.

References

1. Badshah G et al (2014) Mobile cloud computing & mobile battery augmentation techniques: a survey. IEEE Syst J
2. Patti A et al (2011) CloneCloud: elastic execution between mobile device and cloud. In: Proceedings of EuroSys
3. Lee BS et al (2011) µcloud: towards a new paradigm of rich mobile applications. Procedia Comput Sci 5:618–624
4. Niu R et al (2013) An energy efficient multisite offloading algorithm for mobile devices. Int J Distrib Sens Netw
5. Balasubramanian A et al (2010) MAUI: making smart phones last longer with code offload. In: Proceedings of the 8th international conference on mobile systems, applications, and services
6. Wang X et al (2015) Energy and delay tradeoff for application execution in mobile cloud computing. IEEE Syst J 11
7. Nandedkar SC, Mohammad Saad S (2014) Energy efficient mobile cloud computing. Int J Comput Sci Inf Technol 5(6):7837–7840. ISSN: 0975-9646
8. Yang L et al (2012) A framework for partitioning and execution of data stream application in mobile cloud computing. In: Proceedings of IEEE fifth international conference on cloud computing
9. Gani A et al (2015) Energy efficient computational offloading framework for mobile cloud computing. J Grid Comput

drksujatha23@gmail.com

A Bounding Box Approach for Performing Dynamic Optical Character Recognition in MATLAB



Poonam Chaturvedi, Mohit Saxena and Bhavna Sharma

Abstract OCR is used to recognize written or optical generated text by the computer. Machine learning and artificial intelligence are relying frequently on such automation process with high accuracy. This paper present setting of the threshold value is once for whole bounding box algorithm rather than the random threshold value. Region properties of the image measure in the second and final module of our article. In the proposed approach, the final extraction of optical character is done by removing all the feature vectors having pixels less than 30. This process will subsequently increase the accuracy of recognition and visual effects as well. Old and new data sets are implemented by the proposed algorithm. After that, a comparative analysis was done for both outputs of the proposed algorithm. Proposed algorithm extracts different optical characters at the same time so as to reduce time complexity as well.

Keywords MATLAB · Recognition · Visual effects · Threshold
OCR

1 Introduction

Optical character recognition (OCR) is the distinction in physically formed or constructed substance into an electronic course of action, which can be secured, deciphered, and orchestrated by a PC. It can be utilized as a snappy information

P. Chaturvedi · M. Saxena (✉)
Apex Institute of Engineering and Technology, Jaipur, India
e-mail: apexmohit@gmail.com

P. Chaturvedi
e-mail: poonam.chaturvedi26@gmail.com

B. Sharma
Jaipur Engineering College and Research Centre, Jaipur, India
e-mail: bhavnajecrc@gmail.com

© Springer Nature Singapore Pte Ltd. 2019

V. S. Rathore et al. (eds.), *Emerging Trends in Expert Applications and Security*,
Advances in Intelligent Systems and Computing 841,
https://doi.org/10.1007/978-981-13-2285-3_15

drksujatha23@gmail.com

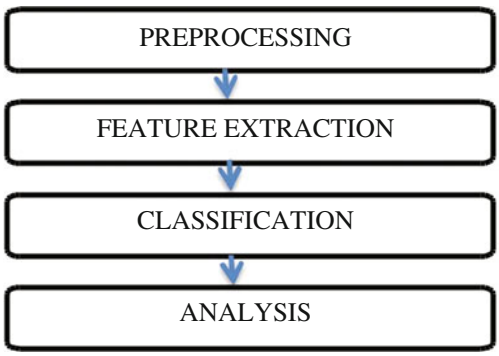


Fig. 1 Steps of OCR [12]

input framework for a front line PC. Any OCR framework depends upon the four key advances (Fig. 1):

- preprocessing
- feature extraction
- classification
- analysis.

1.1 Preprocessing

Pre-getting ready is one of the preliminary walks in character affirmation. Before the rough data is used for incorporate extraction, it needs to encounter assured preliminary methods with the objective for appropriate results. This movement helps in altering the deficiencies in the data which may have happened as a result of the confinements of the sensor. The commitment for a system may be taken in different common conditions. The same inquiry may give apparent pictures when taken in different time and actions. In this we get data that will be straightforward for the system to chip for making definite result.

2 Literature Review

The earlier strategies are nearly spatial primarily based approach. In a spatial domain, the watermark is embedded into the host photo by using directly modifying the pixel values, i.e., only example is to embed the watermark within the least extensive bits of image pixels [1]. Spatial location is straightforward to enforce and requires no particular photo for detection. However, it regularly fails underneath signal processing assaults including filtering and compression and having relative low-bit capability. A simple image cropping operation may additionally do away

drksujatha23@gmail.com

with the optical character recognition (OCR) is the transformation of transcribed or wrote content into an electronic configuration, which can be put away, translated and handled by a PC. It can be utilized as immediate information input strategy for a cutting-edge PC. This segment gives several kinds of virtual watermarking techniques found in the instructional literature. We do now not deliver an exhaustive assessment of the area; however, offer an outline of installed approaches. Max-min calculation [2, 3] was produced to conquer the weakness of the min-min calculation. This calculation is likewise in light of the finish time of the assignment. In this calculation likewise, we initially discover the finishing time of each undertaking on each accessible asset.

3 Problem with the Existing Work

There are various proposed algorithms which perform optical character recognition with online and offline approaches with high and low accuracy of pattern matching. The proposed work targets dynamic approaches which include preprocessing of probe text first to get better accuracy and then analyzes the final result or outputs comes under the recognition task. The existing work has chosen a random and static method, which sometimes mixes data pixel and result in half accuracy and classification error in recognition rate. The proposed algorithm is an advanced approach for overcoming the drawback of the existing work and provides a better accuracy than the existing algorithm. The random approach will require setting the threshold with the random approach but our contemplated approach has a dynamic approach for character recognition in optical manner by using bounding boxes and calculating image region properties. The proposed algorithm is implemented on both new dataset and the older one and then provides the comparitive analysis. Our proposed algorithm extracts the optical characters at the same time so it is as to reduce time complexity as well [1, 4].

Following are the different cases on which comparitive analysis has been performed.

Case I—The existing work has randomly set the maximum threshold for the probe image to extract the optical characters more accurately in the first module of the work. The proposed method contemplated a new technique with a dynamic approach to get the characters on a console window with more accuracy and clarity. Instead of setting the threshold value randomly, we have used bounding boxes with a fixed threshold value for each character [5].

Case II—The existing approach performs segmentation of the objects in the image based on random setting of threshold which sometimes mixes data pixel and results in the classification error in optical recognition. Practically shown how images are convoluted to get smoothen in order to get reduces number of connected feature vector to get better accuracy so maximum number of connected components have been calculated. Proposed algorithm reduces mix data pixel error.

drksujatha23@gmail.com

Case III—The existing work has randomly set the maximum threshold for the probe image to extract the optical characters so it requires more loops to extract image features and optical character but proposed algorithm extract all different character at the same time.

4 Proposed Algorithm

Step1 Take input data set in RGB texture and then perform linearization for making it compatible for the tool.

Step2 Preprocessing of data set should be performed to implement masking and filtering for getting results that are more accurate.

Step3 In order to get accurate image result, image segmentation and feature extraction are performed on input dataset.

Step4 A maximum threshold is set for pixel intensity of 30 so that object lower than this pixel vector can be removed in order to provide better visual effects and less distortion in the dataset and optical character.

Step5 The binarization of image will be repeated to get the final noise-removed image.

Step6 To provide dynamic dimension to our proposed algorithm region attributes of image have been calculated.

Step7 Region properties can be calculated using bounding box for improved optical character recognition.

Step8 Bounding box plotting is done now for getting exact position and edge color of the input data set.

Step9 Now final object extraction will be performed using pseudo code like

```
%% Objects extraction
Figure
For n=1: Ne where L and Ne is label connected component
  [r,c] = find (L==n); r and c are rows and column feature vectors
      n1=imagen(min(r):max(r),min(c):max(c));maximum and minimum feature vectors are
calculated to extract each optical character .
N is the position of optical character and region property attributes.
```

Step10 The above steps will be repeated till the last optical character is not extracted efficiently.

Step11 exit.

drksujatha23@gmail.com

5 Experimental Results

Figure 2 [6] is just used as an input data set which depicts the car number plate and the first module will load this into the tool for preprocessing then the next process will be implemented. The given image is used in implementing the existing work with random approach of setting the threshold value.

Figure 3 shown above is just the linearization of the input dataset so that it becomes compatible to the tool to get processed further. The 2-dimensional one is only acceptable in the tool for getting the number of maximum connected components so that we can get that number reduce in the proposed method to get the accuracy improved along with better time complexity and better extraction capabilities. The existing work has randomly set the maximum threshold for the probe image to extract the optical characters more accurately in the first module of the work.

Proposed algorithm contemplated a new technique with dynamic approach to get the characters on a console window with more accuracy and clarity. Instead of setting the threshold value randomly, we have used bounding boxes in the work.

Figure 4 is showing the segmentation of different optical characters so that the content can be recognized accurately but the existing work has a drawback that it recognizes a single character for each time.

Figure 5 has depicted the extracted optical character, which has been performed by using existing work algorithm, so we can see that only single optical character can be recognized at the single time this was one of the drawback of existing work.



Fig. 2 Input dataset of car number plate



Fig. 3 Binarization of the input data set

drksujatha23@gmail.com



Fig. 4 Optical segmentation of the probe image to get content separated



Fig. 5 The single optical extracted character by using existing algorithm

6 Comparison Between Existing and Proposed Algorithm

Existing approach	Proposed approach
1. Perform segmentation of the object in an image based on random setting of threshold which sometimes mixes data pixel and result in classification error in OCR	1. In this approach, we consider different threshold values for each bounding box and improve the accuracy of OCR and visual effects
2. Randomly set the maximum threshold for the probe image to extract the optical characters more accurately in the first Module	2. Dynamic approach get the characters on a console window with more Accuracy and clarity

7 Conclusion

There are various algorithms proposed for optical character recognition in the computer vision field. It is a desired topic in machine learning and artificial intelligence field. The previous work has referenced and was totally a static approach in which threshold and more character text are not recognized at the same time. Our algorithm is based on recognition oriented which also reduces the time complexity of overall operations. The existing work has randomly set the maximum threshold for the probe image to extract the optical characters more accurately in the first module of the work. New technique with dynamic approach to get the characters on

drksujatha23@gmail.com

a console window with more accuracy and clarity. Bounding box overlapping problem was also detected in this algorithm.

The existing method is based on median filter for noise removal. It basically finds the edges existing in the probe image and then preprocess the content. Practically, images are convoluted to get smoothen and reduces the number of the connected feature vector to get better accuracy for that we have calculated the maximum number of connected components. The future scope of our proposed also can be in online character recognition system what we have done is an offline process but the approach is dynamic.

References

1. Chaudhuri A et al (2017) Optical character recognition systems for different languages with soft computing. Stud Fuzziness Soft Comput 352. https://doi.org/10.1007/978-3-319-50252-6_2, Springer International Publishing AG
2. Cho S-B (1997) Neural-network classifiers for recognizing totally unconstrained handwritten numerals. IEEE Trans Neural Netw 8(1), January 1997
3. Chaudhuri BB, Pal U, Mitra M Automatic recognition of printed oriya script
4. Basa D, Meher S (2011) Handwritten Odia character recognition. In: Presented in the national conference on recent advances in microwave tubes, devices and communication systems, Jagannath Gupta Institute of Engineering and Technology, Jaipur, March 4–5 2011.
5. Plamondon R, Srihari SN (2000) On-line and off-line handwritting character recognition: a comprehensive survey. 1EEE Trans Pattern Anal Mach Intell 22(1)
6. Vamvakas G, Gatos B, Stamatopoulos N, Perantonis SJ (2016) A complete optical character recognition methodology for historical documents. In: The eighth IAPR workshop on document analysis systems, January 4–5
7. Vamvakas G, Gatos B, Stathopoulos N, Perantonis SJ (2008) A complete optical character recognition methodology for historical documents. In: The eighth IAPR workshop on document analysis systems
8. Kimura F, Wakabayashi T, Miyake Y (1996) On feature extraction for limited class problem, August 25–29
9. Deshmukh S, Ragha L (2009) Analysis of directional features—stroke and contour for handwritten character recognition. In: 2009 ieee international advance computing conference (IACC 2009) Patiala, India, pp 6–7, March 2009
10. Blumenstein M, Verma BK, Basli H (2003) A novel feature extraction technique for the recognition of segmented handwritten characters In: 7th international conference on document analysis and recognition (ICDAR'03) Edinburgh, Scotland, pp 137–141
11. Bhowmik T, Parui SK, Bhattacharya U, Shaw B An HMM based recognition scheme for handwritten Oriya numerals
12. Pai N, Kolkure VS Optical character recognition: an encompassing review. IJRET: Int J Res Eng Technol eISSN: 2319-1163 | pISSN: 2321-7308

drksujatha23@gmail.com

Performance Comparison of LANMAR and AODV in Heterogenous Wireless Ad-hoc Network



Madhavi Dhingra, S. C. Jain and Rakesh Singh Jadon

Abstract Wireless ad-hoc network is a self-configuring network which works without the help of centralised devices. Dynamic nature of network topology is a major concern in this kind of network and thus it requires efficient routing protocols. There are various routing protocols which are working in this area for specific objectives. The performance of each routing protocol varies with network configuration and tits specific working. This paper has focused on the LANMAR and the AODV routing protocol and compared their performance on various parameters in heterogenous networks.

Keywords We LANMAR · AODV · Routing protocols · MANET

1 Introduction

This instruction file Wireless ad-hoc network is infrastructure less network where no routers are fixed. All nodes are mobile and capable of moving in the network. Nodes act as routers, identify and maintain the routing inside the network. These networks are inexpensive in comparison to infrastructure-based network. There are various MANET routing protocols having distinct characteristics. These protocols are classified on the basis of several points that include

M. Dhingra (✉) · S. C. Jain
Amity University Madhya Pradesh, Gwalior, Madhya Pradesh, India
e-mail: madhavi.dhingra@gmail.com

S. C. Jain
e-mail: scjain@gwa.amity.edu

R. S. Jadon
MITS, Gwalior, India
e-mail: rsjadon@gmail.com

© Springer Nature Singapore Pte Ltd. 2019

V. S. Rathore et al. (eds.), *Emerging Trends in Expert Applications and Security*,
Advances in Intelligent Systems and Computing 841,
https://doi.org/10.1007/978-981-13-2285-3_16

drksujatha23@gmail.com

- Communication model
- Structure
- State Information
- Cast Property
- Scheduling

One of the important classifications of routing protocols is casting property, based on which it is divided into two types: Unicast and Multicast Routing Protocol. Unicast protocol sends the messages to a single node or host in the network while multicast protocol sends the messages to a group of hosts of the network.

There are various unicasting routing protocols LANMAR and AODV are the unicast routing protocols which are discussed in this paper for performance comparison. The main work of this paper is to assess these protocols in the simulated environment. The results are evaluated on Qualnet Simulator.

2 Working Principle of AODV and LANMAR

2.1 AODV Routing Protocol

The Ad-hoc On-Demand Distance Vector algorithm allows dynamic multiple node routing. This protocol can work small as well as large sized networks [1]. The AODV protocol can be used in only those cases where two nodes do not have a specific path between each other. A precursor list is maintained to keep track of the IP addresses of the neighbouring nodes. This information is used in the routing table. This algorithm does not maintain routes of inactive nodes. This protocol enables the mobile nodes to inform the other nodes that are to be affected by link breakages and change in topology. To perform the above functions, this protocol defines three types of messages. First, Route Requests (RREQs) messages initiate the route finding process. Second, Route Replies (RREP) messages finalise the path and third, Route Errors (RERRs) message a inform the nodes of the network about the breakage of the link used in an active route.

3 LANMAR Routing Protocol

LANMAR is basically designed for scalable networks. Landmark routing was first discovered in wired networks [2]. This method requires multiple levels of hierarchical addressing that has to be predefined. The hierarchical address specifies the location of the node and helps in determining its path. Each node has complete knowledge about the path of all other nodes in the network and the landmarks at different levels in the hierarchy. All the paths are defined in top-down manner.

drksujatha23@gmail.com

LANMAR uses the concept of landmark and scoped routing. Its wired use is extended in the wireless ad-hoc network. This scheme does not need the predefined addresses of the hierarchy. Instead, it makes use of the specified landmarks to have knowledge about the different networks. Landmarks are basically a group of nodes having same characteristics and goal. The information about the landmark routes is sent all over the network by using distance vector procedure. The route to a landmark is sent throughout the network by following the process of distance vector algorithm.

All the landmarks store the routing information in the summarised form, thereby reducing the size of the routing table. These landmarks are also aware of the change of routing in case of traffic overhead. Thus, this protocol provides an efficient and scalable environment for ad-hoc networks. This algorithm provides more advantages in terms of reducing the routing packet size and updating the correct routes to mobile nodes. It achieves high data packet delivery ratio.

4 Proposed Methodology and Simulation Setup

The methodology [3] used in simulator is divided into following parts

1. The scenarios is set up to create the heterogenous network of 14 nodes. The wired and wireless network both are set up using the nodes and the wired and wireless links. CBR is given between nodes.
2. The source to destination links are made through the heterogenous network.
3. In the third step, routing protocol is specified at the MAC layer and simulation is run.
4. The analysis reports are seen and compared for evaluating the performance.

5 Node Placement Scenario

The heterogenous network is created by creating two hierarchies of wired nodes and one of wireless system (Fig. 1).

5.1 Simulation Setup

The simulation of the heterogenous mobile ad-hoc network is done on Qualnet [4]. The performance of wireless network is evaluated by applying routing protocols on hierarchical network [5–8]. Wired and wireless hierarchy are set up and Constant Bit Rate (CBR) [9, 10] is applied between different nodes. The scenario size is set as 1000 * 1000 sq with 14 nodes. Packet size of 512 bytes is used. Wired link is given from {2} to {9} and from {8} to {14} (Table 1).

drksujatha23@gmail.com

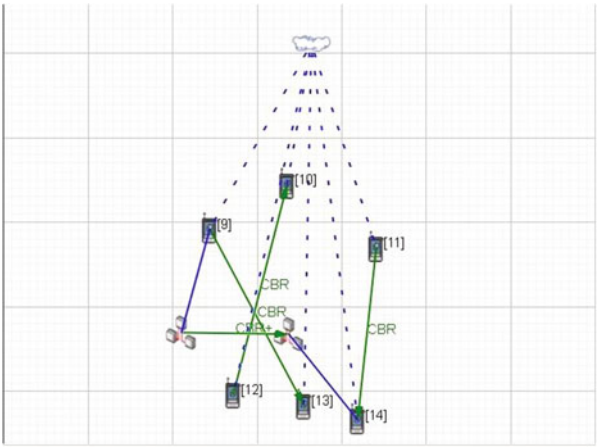


Fig. 1 Node placement scenario of heterogenous wireless MANET

Table 1 Simulation parameters

Parameters	Values
Number of nodes	14
Simulation area	1000 * 1000
Routing protocols	AODV/LANMAR
Packet size	512 bytes
Traffic Type	Constant Bit Rate (CBR)
Hierarchy 1	{2, 9}
Hierarchy 2	{8, 14}
Wireless subnet	(9 through 14}
Wired subnet	{1 through 4}, {5 through 8}
Simulation time	200 s

6 Discussion and Simulation Result

The performance of AODV and LANMAR protocols is evaluated based on the following parameters:

1. Average unicast end to end delay (Figs. 2 and 3)
2. Average unicast jitter (Figs. 4 and 5)
3. Unicast Received Throughput (Figs. 6 and 7)
4. Unicast packets sent to and received from channel 802.11 MAC (Figs. 8, 9, 10 and 11 and Table 2)

drksujatha23@gmail.com

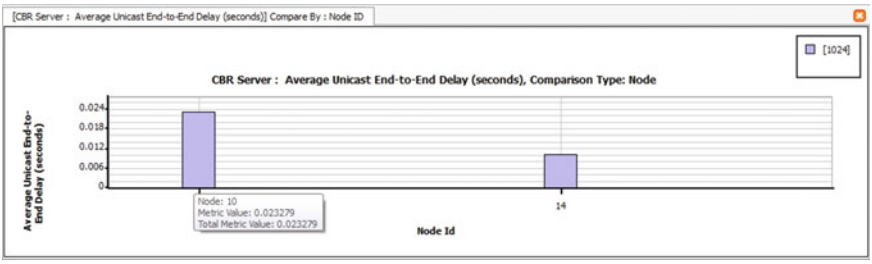


Fig. 2 AODV: CBR Server

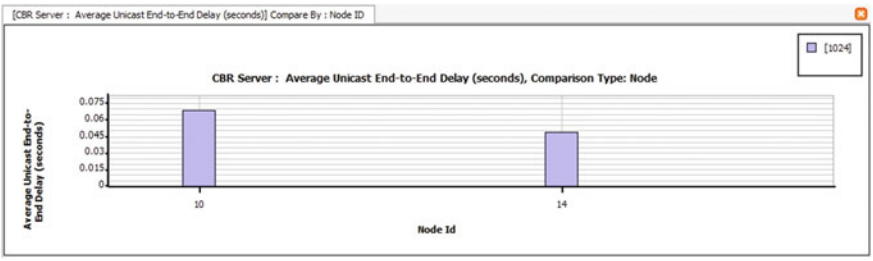


Fig. 3 LANMAR: CBR Server

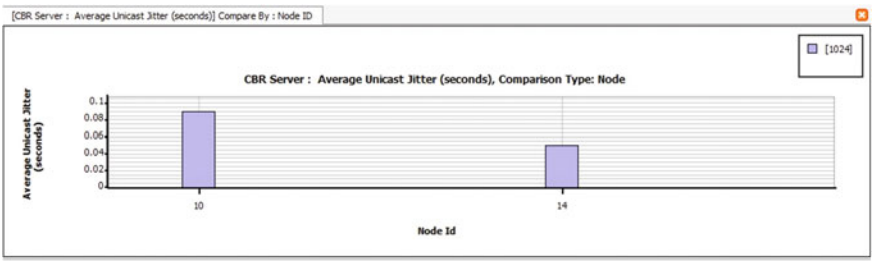


Fig. 4 AODV: CBR Server

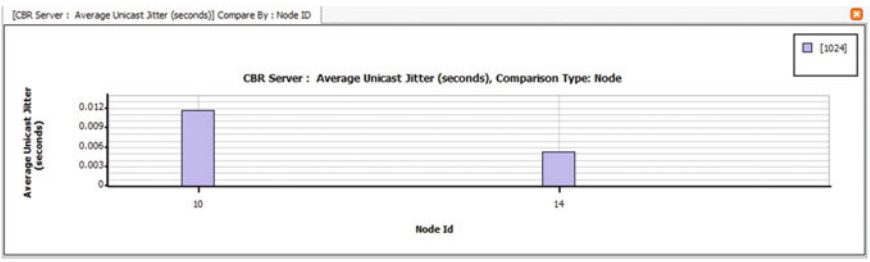


Fig. 5 LANMAR: CBR Server

drksujatha23@gmail.com

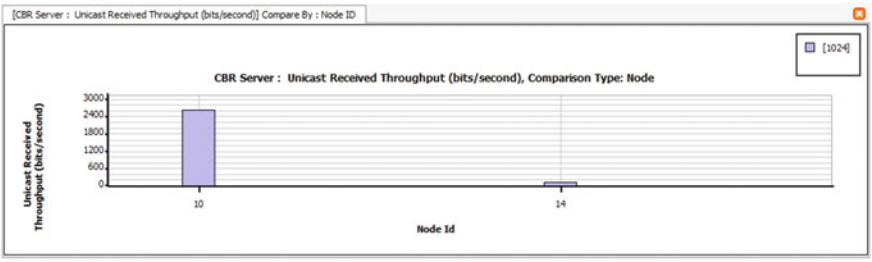


Fig. 6 AODV: CBR Server

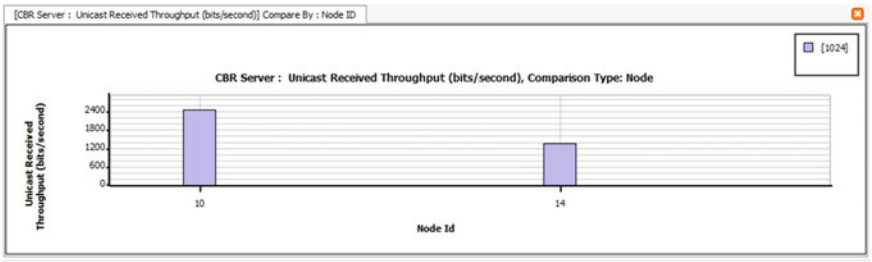


Fig. 7 LANMAR: CBR Server

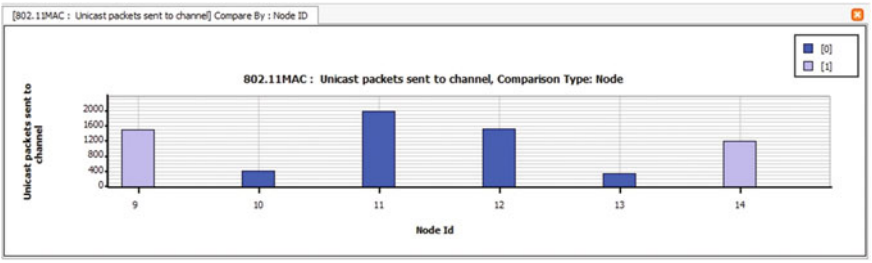


Fig. 8 AODV 802.11 MAC unicast packets sent

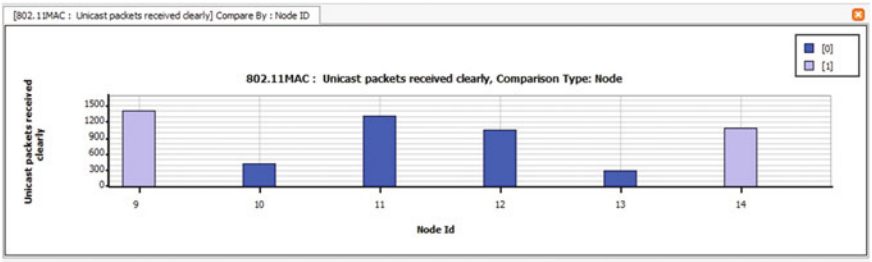


Fig. 9 AODV 802.11 MAC unicast packets received

drksujatha23@gmail.com

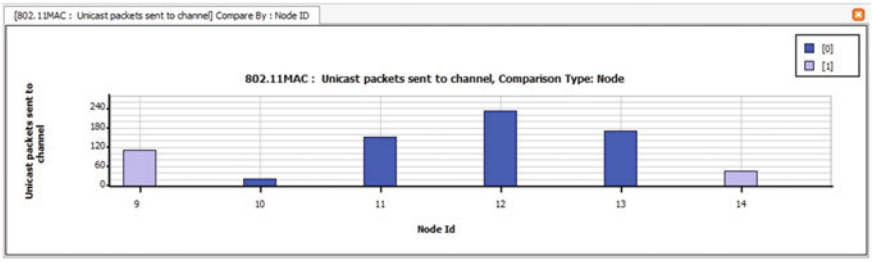


Fig. 10 LANMAR: 802.11 MAC unicast packets sent

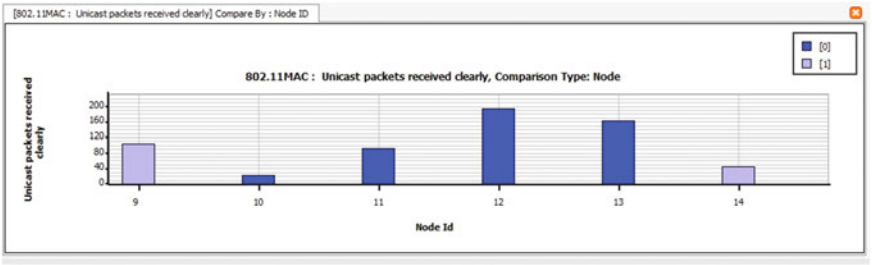


Fig. 11 LANMAR: 802.11 MAC unicast packets received

Table 2 Simulation results

Parameters	AODV values	LANMAR values
Average unicast end to end delay (in seconds)	0.065 at node {10}, 0.047 at {14}	0.023 at {10}, 0.010 at {10}
Average jitter	0.082 at {10}, 0.05 at {14}	0.012 at {10}, 0.005 at {14}
Unicast received throughput (bits/sec)	2400 at {10}, 1400 at {14}	2600 at {10}, 100 at {14}

7 Conclusion

This paper has analysed the performance of AODV and LANMAR routing protocol based on some parameters by using Qualnet simulation tool. In heterogenous wireless networks, LANMAR routing protocol is best in case of unicast end to end delay, average jitter. AODV and LANMAR both protocols give the optimal throughput. Packet Dropping ratio of LANMAR is very minimal. Overall, LANMAR is best suited for large-scale mobile ad-hoc wireless heterogenous network.

drksujatha23@gmail.com

References

1. Perkins et al (2003) Ad hoc on-demand distance vector (AODV) routing, RFC 3561, July 2003
2. Tsuchiya PF (1988) The landmark hierarchy: a new hierarchy for routing in very large networks. Comput Commun Rev 18(4):35–42
3. Study and performance analysis of routing protocol based on CBR, 1877-0509 © 2016 Published by Elsevier. Procedia Comput Sci 85:23–30 (2016)
4. Kumara J, Singh A, Pandac MK, Bhadauriad HS. QualNet documentation, QualNet 5.0 Model Library: Advanced Wireless. http://www.scalablenetworks.com/products/Qualnet/download.php#docs
5. Elizabeth R, Toh CK (1999) A review of current routing protocols for ad hoc mobile wireless networks: RFC 2409. IEEE Pers Commun
6. Perkins CE, Royer EM, Das SR (2002) Ad hoc on-demand distance vector (AODV) routing, Internet Draft, draft-ietf-manet-aodv-10.txt, work in progress
7. Rath SR (2009) Study of performance of routing protocols for mobile Adhoc networking in NS-2, NIT, Rourkela
8. Nand P, Sharma SC (2011) Performance study of broadcast based mobile Adhoc routing protocols AODV, DSR and DYMO. Int J Secur Appl 5(1)
9. Siva Rammurty C, Manoj BS (2011) Ad hoc wireless networks architectures and protocols. ISBN 978-81-317-0688-6
10. Johnson DB, Maltz DA. Dynamic source routing in ad hoc wireless networks, Computer Science Department, Carnegie Mellon University, Avenue Pittsburgh, PA 15213-3891

drksujatha23@gmail.com

An Efficient Approach for Power Aware Routing Protocol for MANETs Using Genetic Algorithm



Renu Choudhary and Pankaj Kumar Sharma

Abstract Mobile Adhoc Networks MANETs are very popular networks which are in use now a day. These are infrastructure-less networks. The remaining battery power of nodes is an important resource in MANETs. Routing process is a very power consuming process in MANETs. If a node involves as an intermediate node in transferring the data, then its remaining battery power decreases rapidly and the node may die (stop working) because of lack of battery power. So routing protocols should be designed so that the remaining battery power of nodes should be used efficiently. It will result in increased network lifetime. This paper proposed a new power-aware routing protocol for MANETs. The proposed algorithm uses Genetic Algorithm to find a path that consumes minimum power while routing data in MANETs. The algorithm is implemented on a sample network in JAVA programming and power consumption is reduced in routing data.

1 Introduction

Wireless networks are very popular networks which use wireless communication. These are dynamic networks in which nodes of the network may change their location with time. These networks are of two types (a) Infrastructure networks (b) Infrastructure-less networks.

Infrastructure networks—These types of networks need infrastructure for their networks. Wireless LAN such as Bluetooth, Wi-Fi, and cellular networks are infrastructure networks.

Infrastructure-less networks—In these networks, no infrastructure is available but terminals are fit. The nodes of the network communicate with each other

R. Choudhary (✉) · P. K. Sharma
Department of Computer Science, Government Women Engineering College, Ajmer (GWECA), Ajmer, Rajasthan, India
e-mail: rennu1992@gmail.com

P. K. Sharma
e-mail: pankaj.gmeca@gmail.com

© Springer Nature Singapore Pte Ltd. 2019

V. S. Rathore et al. (eds.), *Emerging Trends in Expert Applications and Security*, Advances in Intelligent Systems and Computing 841,
https://doi.org/10.1007/978-981-13-2285-3_17

drksujatha23@gmail.com

without any infrastructure. Some examples are emergency services, sensor networks, disaster recovery networks, etc.

In MANETS, power is consumed to broadcast the messages for route request. So the power level of nods is an important issue in MANETs. If the battery of nodes goes down, then that node cannot communicate with other nodes.

Because of battery operated nodes, the routing algorithm must be designed to minimize the power consumption of nodes in the routing process. The routing algorithms which are designed to consider the power consumption are classified as power aware routing protocols in MANETs. In this paper, a new power-aware routing protocol is proposed using Genetic Algorithm. Genetic Algorithm is an optimization algorithm which can optimize the power consumption of nodes in routing data in MANETs. Genetic Algorithm works on its special operators such as selection, crossover, and mutation.

Section 2 of this paper includes the literature survey and highlight the recent research done in this field. Section 3 includes the proposed work of this paper. Sections 4 and 5 are explaining the results part and conclusion part of this paper.

2 Related Work

Safa et al. [1] proposed a power-aware routing protocol for MANETs. The protocol is avoiding the nodes which are exhausted in data transmission. The algorithm avoids such nodes to be used again and again for routing purpose. Bheemalingaiah et al. proposed [2] a power-aware multipath source routing in mobile Adhoc networks. Bhople and Waghmare [3] proposed routing protocol for MANET with power utilization optimization. The algorithm selects a node that is having the highest remaining battery power. The algorithm also consumes less power in transferring power. Singh et al. [4] proposed location aided energy aware routing protocol for MANET. It uses linear regression and curve intersection for its power aware algorithm. Manohari and Ray proposed [5] an energy efficient multipath routing protocol for MANET. The algorithm uses current battery status and current traffic for power aware routing protocol. Pawan et al. [6] use cluster head to propose routing protocol for MANETs using cluster head. In the paper, the author nominates a node in the adhoc as the cluster head. All the communication to the other Adhoc is done via that cluster head.

3 Proposed Work

In this paper, a new approach for power-aware routing protocols is proposed. The new routing algorithm is based on table-driven protocols in which every node maintains a table that stores the information about paths to transfer data to all neighboring nodes in the network. Because MANETs are dynamic networks, i.e.,

drksujatha23@gmail.com

the location of a node may change with time so every node has to update the routing table. The new routing protocol uses this routing table to get the information about different nodes and then use Genetic Algorithm to find a path to transfer data from current node to the destination node. Genetic Algorithm is a soft computing technique which is used to solve optimization problems. In routing, a node is selected in routing path in such a way that it considers Remaining Battery Power (RBP) of that node while selecting that node as an intermediate node in a path. Further, the fitness of a path is dependent on two factors (1) Average Remaining Battery Power of all the nodes in the path (2) Number of nodes selected in that path.

To solve this problem using genetic algorithm, the fitness of a path is dependent on two things, i.e., average remaining battery power of nodes and number of nodes in that path. The genetic algorithm will discard those paths which are less fit, i.e., those paths which are selecting nodes with less remaining battery power or paths in which more number of nodes are selected.

Algorithm 1 is showing the steps of the proposed genetic algorithm for power-aware routing protocol.

Algorithm 1 : Proposed Power Aware Routing using Genetic Algorithm
Input : Network Information, Source Node, Destination Node
Output : Routing Path

1. Generate a Random Initial population of routing paths
2. Calculate fitness of every path in the population using proposed algorithm by considering average remaining battery power and number of nodes in the path (explained later in this section)
3. Repeat steps 4 to 8 till an optimal solution is not reached
4. Select a set of parents from initial population to perform cross over
5. Perform cross over to generate children and calculate fitness of every children
6. Add newly generated children paths in population and sort population by fitness
7. Remove those paths from population which are less fit
8. Perform mutation and generate population for next iteration.

Genetic parameters used in this proposed genetic algorithm are as follows:

Initial Population Generation technique:	Random initial population generation
Crossover Operator Used:	One-point crossover
Crossover Rate:	40%
Mutation Rate:	2%

The proposed Algorithm 2 is implemented on a sample network of 38 nodes. Figure 1 is showing a sample of MANET of 38 nodes. Figure 1 is showing a MANET which was used in [6] as a sample network. Here, three types of nodes are available which are Cluster Head (CH), Normal Node (NN), or Gateway Node (GN). In this paper, all nodes are treated equally. The given MANET is simulated in JAVA programming language using NetBeans IDE 8.2 and JDK 1.8. Figure 2 is showing a snapshot of actual implementation.

Novelty in this work—This work design a new formula to calculate the fitness of paths of the population. It considers two factors which are (1) Number of nodes in

drksujatha23@gmail.com

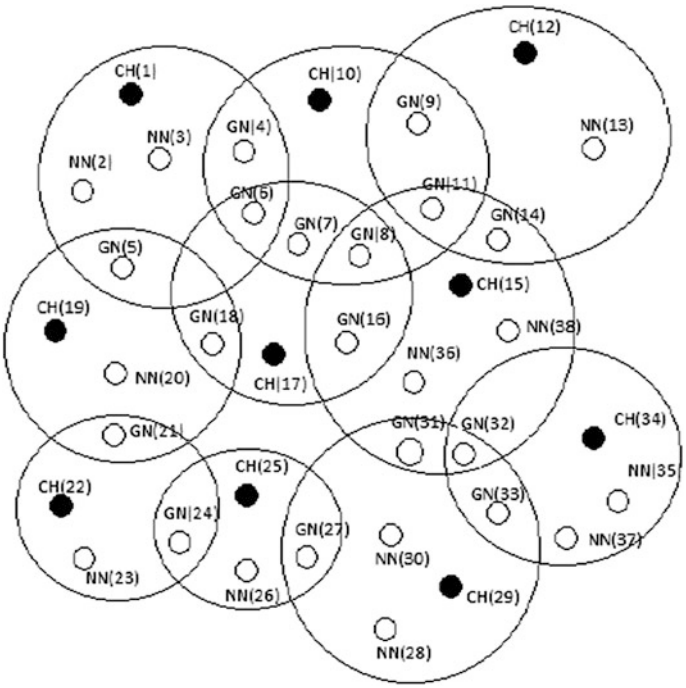


Fig. 1 A sample MANET of 38 nodes and 10 Adhoc

the path (2) Average remaining battery power of nodes in the path. The formula is designed to (1) Reduce the number of nodes in the path and (2) Use nodes with high remaining battery power.

(1) Number of nodes in the path—Paths with more number of nodes are less fit as compared to paths with less number of nodes. If a path is having more than the communication will consume battery power of more nodes which will reduce the overall lifetime of the network. That is why the number of nodes in the path should be less.
(2) Remaining battery power of nodes in the path—This algorithm selects those nodes for routing which are having comparative high value of remaining battery power.

As the average remaining battery power of nodes in a path increases, its fitness also increases. As the number of nodes in the path increases, the fitness of a path decreases. Thus, a path is considered as fit/good if it has high value of AVG_RBP and low value of number of nodes.

In Fig. 2, a snapshot of simulation is shown. Here, node-2 is selected as a source node and node 3, 5 is selected as a destination node. The population is showing 20 paths starting from path-0 to path 19. Here for every path the average remaining battery power, number of nodes, fitness value, and list of cities in the path are

drksujatha23@gmail.com

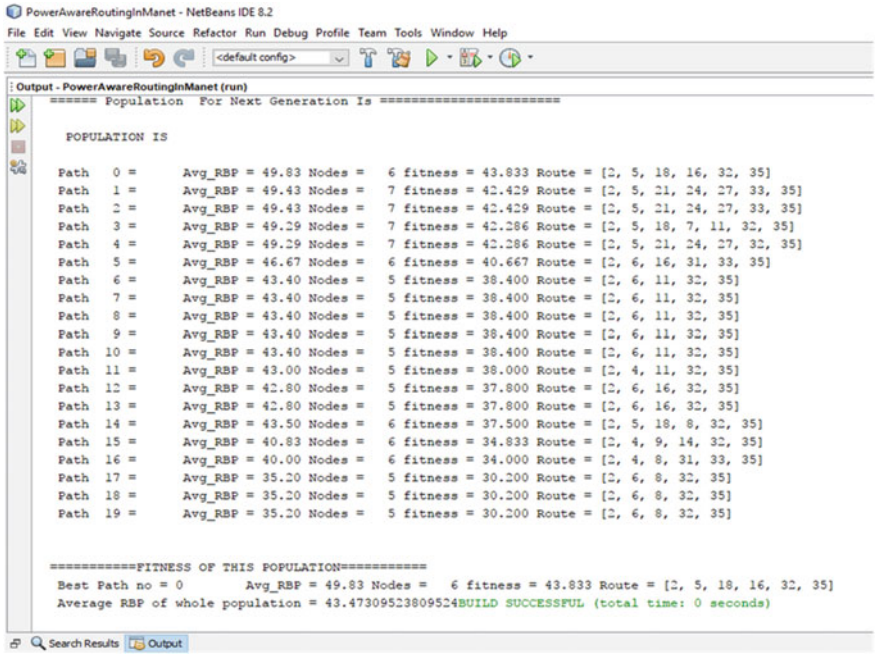


Fig. 2 A snapshot of simulation result

shown. The population is sorted by decreasing the value of the fitness. The best path is shown at path number 0, i.e., the first path.

Best Path of the population
Path No = 0
Avg_RBP = 49.83
Number of active nodes = 6
Fitness = 43.833
Route = [2, 5, 18, 16, 32, 35]

The best path is having average remaining battery power (AVG_RBP) equal to 49.83, number of nodes in the path equal to 6, fitness equal to 43.83 and a list of nodes in the path, i.e., [2, 5, 18, 16, 32, 35]. The fitness of a path is dependent on average remaining battery power (AVG_RBP) and number of nodes. Here, fitness of a path = Avg_RBP − number of nodes

$$\text{Hence for best path fitness} = 49.83 = 6 = 43.83$$

Thus the proposed algorithm selects a path with less number of nodes and high value of remaining battery power.

drksujatha23@gmail.com

4 Result Analysis

In the previous work [6], the author simulated the same MANET and discover a path having number of nodes equal to 9 [2-1-6-10-8-15-32-34-35]. Using this proposed algorithm, a path having number of node equal to 6 is discovered. Thus, the proposed algorithm is able to discover and find a path which is 33% better, i.e., having 33% less number of active nodes in a path. Figure 3 is showing a graph comparing number of nodes in the path using existing algorithm [6] and proposed algorithm.

5 Conclusion and Future Scope

The power-aware routing protocols are proposed to reduce the power consumption in routing. This paper a power-aware routing protocol using a genetic algorithm is proposed. The simulation results show that the proposed algorithm is 33% better than the existing algorithm. In future, the same algorithm is to be applied and tested on MANETs having thousands of nodes. Further, the formula to calculate the fitness of a path can be adjusted and rewritten in future. In future, work can also be done to reduce the complexity of the algorithm for the proposed genetic algorithm.

References

1. Safa H, Karam M, Moussa B (2013) A novel power aware heterogeneous routing protocol for MANETs. In: 2013 IEEE 27th international conference on advanced information networking and applications. https://doi.org/10.1109/aina.2013.36
2. Bheemalingaiah M et al (2017) Performance analysis of power-aware node-disjoint multipath source routing in mobile ad hoc networks. In: 2017 IEEE 7th international advance computing conference
3. Bhople NB, Waghmare JM (2016) Energy routing protocol with power utilization optimization in MANET. In: IEEE international conference on recent trends in electronics information communication technology, 20–21 May 2016, India. 978-1-5090-0774-5/16/ $31.00 © 2016 IEEE 1371
4. Singh K, Sharma A (2015) Linear regression based energy aware location-aided routing protocol for mobile ad-hoc networks. In: 2015 International conference on computational intelligence and communication networks. 978-1-5090-0076-0/15 $31.00 © 2015 IEEE. https://doi.org/10.1109/cicn.2015.30
5. Manohari PK et al (2015) EAOMDV: an energy efficient multipath routing protocol for MANET. In: 2015 IEEE power, communication and information technology conference (PCITC), Siksha 'O' Anusandhan University, Bhubaneswar, India. 978-1-4799-7455-9/15/ $31.00 © 2015 IEEE
6. Pawan et al (2016) An efficient power aware routing protocol for mobile adhoc networks using cluster head. In: 2016 International conference on computing for sustainable global development (INDIACom). 978-9-3805-4421-2/16/$31.00 @ 2016 IEEE

drksujatha23@gmail.com

Multi-purposed Question Answer Generator with Natural Language Processing



Hiral Desai, Mohammed Firdos Alam Sheikh and Satyendra K. Sharma

Abstract Artificial Intelligence is a way of making a computer that works in a similar manner that the smart human think. Machine language, which is a type of artificial intelligence that provides computers to learn without being explicitly programmed or ruled. A Multi-Purposed Question Answer Generator which is based on AI in which machine automatically generates the question from contests and also give their answers. So, it is very useful for teachers for giving an assignment and their solution to students. It is also useful for giving one mark question as well its answer to students after every chapter. So, it eliminates the tedious job of teachers and gives an easy solution.

Keywords Artificial intelligent · Human brain · Machine language
Train data · NLP

1 Introduction

Machine learning emphasizes on the generation of computer programs that can teach themselves to fatten and alter when disclosure to new data. Machine learning is a way of teaching computers to accomplish and enhance prediction based on some data [1].

H. Desai · M. F. A. Sheikh (✉)
Pacific School of Engineering, Surat, India
e-mail: firdos.sheikh@gmail.com

H. Desai
e-mail: hiral8.desai@yahoo.com

S. K. Sharma
MITRC, Alwar, India
e-mail: skpacific323@gmail.com

© Springer Nature Singapore Pte Ltd. 2019

V. S. Rathore et al. (eds.), *Emerging Trends in Expert Applications and Security*,
Advances in Intelligent Systems and Computing 841,
https://doi.org/10.1007/978-981-13-2285-3_18

drksujatha23@gmail.com

Machine learning tasks are analyzed in three categories

(1) Supervised Learning
(2) Unsupervised Learning
(3) Reinforcement Learning.

(1) **Supervised Learning**

In supervised learning, we will give some example inputs and pretended output to computer and the intention is mapping input to pretended output with general rule. Examples of supervised learning are as follows.

You are a kid, you see different types of vehicles on roads, your guardian tells you that this particular vehicle is a car…after him/her giving you tips a few times, you see a new type of car that you never saw before—you identify it as a car and not as an auto or a truck or a bike. You identify because you have a teacher to guide you and learn concepts, such that when a new sample comes your way that you have not seen before, you may still be able to identify it.

Assume we are on a road and our task is to arrange the same types of vehicles at one area. Suppose the vehicles are bike, car, auto.

Suppose the vehicles are bike, car, auto. So we know from our previous work that the shape and features of every vehicles so it is easy to manage the same type of vehicles at one area. Here, your previous work is called as train data in data mining. So we learn the things from our train data. This is because of we have a response variable which says you that if some vehicles have so and so features it is auto, like that for every vehicles.

- This type of data you will get from the train data.
- This concept of learning is called as supervised learning.
- This problem solving come under classification [5].

(2) **Unsupervised Learning**

In this learning algorithm, no label is given to find a structure, leaving it in its input. The structure can be derived by grouping of the data that are based on the relationship between the variables in the data. There is no feedback in unsupervised learning on prediction results.

Suppose you are on a road and your task is to manage the same types of vehicles at one area. This time you do not know anything about that vehicles, you are first time seeing this vehicle so how can you manage the same type of vehicles. What will be the first step? you can choose one vehicle and you will choose any physical character of that particular vehicle. Assume you taken the number of wheels. Then you will arrange them base on the number of wheels, then the groups will be something like this.

TWO WHEELS GROUP: Cycle and Bike.
THREE WHEELS GROUP: Auto.
FOUR WHEELS GROUP: Car and Truck.

drksujatha23@gmail.com

So now you will take another physical character as size, so now the groups will be something like this.

TWO WHEELS AND BIG SIZE: Bike.
TWO WHEELS AND SMALL SIZE: Cycle.
FOUR WHEELS AND BIG SIZE: Truck and Tempo.
GREEN COLOR AND SMALL SIZE: Car.

Here, you did not know learn anything before means no train data and no response variable. This type of learning is known as unsupervised learning [5].

(3) **Reinforcement Learning**

A program which interacts with a changing environment and it should perform a specific goal, without any teacher explicitly telling it whether it has come near to its goal.

- Game playing. The player knows whether it won or lost, but not how it should have moved.
- Control: A traffic system can measure the delay of cars, but not know how to decrease it.
- Robot path planning: Can measure the actual distance traveled.

2 Application of Machine Learning

Example of Machine language is Microsoft cognitive services (Emotion Detection). In Microsoft, cognitive services machine takes input as image and detects the person's behavior—whether she or he is happy, sad, surprise, and so on (Fig. 1).

There are many applications are available which are based on Machine Learning like NLP Sentiment Analysis (Fig. 2).

In NLP sentiment analysis, machine takes input from user and output as whether the statement is positive or negative [3].



Fig. 1 Emotion detection

drksujatha23@gmail.com

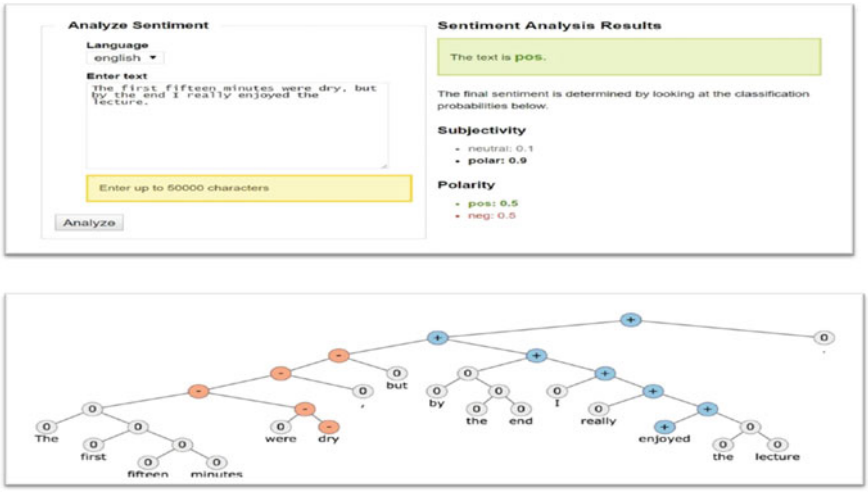


Fig. 2 Sentiment analysis

3 Proposed Multi-purposed Automated Question and Answer Generator Based on Machine Learning

Multi-purposed automated question answer generator is based on supervision or supervised learning. In this concept, we provide contents (input) to machine and machine will automatically generate questions from contents and machine also give the answer to that question automatically. For that, we defined or learn some rules to the machine so, the machine can generate question automatically and give an answer to that question automatically.

An Automatic Question Generator in which machine automatically generates the question from contests so, it is useful for teachers to prepare a assignment or exercise for students. But in this approach of learning to get an answer of particular question, teachers have to refer the whole content so, it is a big task for teachers to prepare a solution of that assignments or exercises and its a time-consuming process [2].

So, to avoid such a problem we design a—"Multi-purposed Automated question and answer generator". In which we design some rules and through rules, we learn machines. Machines take inputs from the user and based on that contents, machine will automatically generate the question from input content and also give the answer of that question at same time. Machine can also find data if we search for particular questions. So, it reduces the time of teachers to make question bank for students and make solution of the same. Through this, teachers can also set a question paper of exams. Through this, one can also get a quiz questions and answer as well. So, with the uses of multi-purposed automated question and answer

drksujatha23@gmail.com

generator peoples can save their time, without reading anything, they are able to get questions and answers as well.

We Solve problems by following levels:

(1) First We work with Statements
(2) Next, we work with Multiple statements
(3) Next, we work with Paragraphs
(4) And so on.

4 Sample Data Given to Machine and Its Desired Output

4.1 Sample Design

Consider the following contents and based on this content, machine automatically generates questions and gives the answer to that question automatically.

Input

There are basically three types of programming languages. (1) machine level language, (2) assembly level language and (3) High level language.

(1) Machine level language

    Machine level programs only understand binary codes. Programs are in form of 0's and 1's. programs are not portable. Not required translator to translate the program because computer directly understand this language.

(2) Assembly level language

    Assembly level programs are not in the forms of 0's and 1's so computer cannot understand this language directly. Assembler tool is used for translation. Programs are not portable.

(3) High level language

    High level programs are not in forms of 0's and 1's so computer cannot understand this language directly. Compiler tool is used for translation. Programs are portable.

Output

Q-1 How many programming languages are there? Ans. There are three programming languages.
Q-2 In which languages programs are not portable?
Ans. Machine level and Assembly level programs are not portable.

Above is the sample question answer. The machine will generate as possible questions from contents and also give the answers.

drksujatha23@gmail.com

4.2 Research Design

Figure shows the Research Design of the current study problem (Fig. 3).

There are basically two main steps of design

(1) We train a machine learning model using our existing labeled data. Labeled data is a data which has been labeled with the outcome which in the case of types of vehicles example whether the vehicle is an auto or car. This is called—model training because the model is learning the relationship between the attributes of the data and the outcome. These attributes might include the number of wheels, shape, weight, etc.
(2) We make predictions on new data for which we do not know the true outcomes. In other words, when a new fruit arrives, we want our train model to accurately predict whether the fruit is an apple or orange without a human examine it.

We want to build machine learning model that accurately predict the labels of our future vehicles, rather than accurately predicting the labels of vehicles we have already received [4].

4.3 Tool to Be Used

(1) SciKit tool
(2) Tensor Flow tool
(3) Natural language toolkit (NLTK).

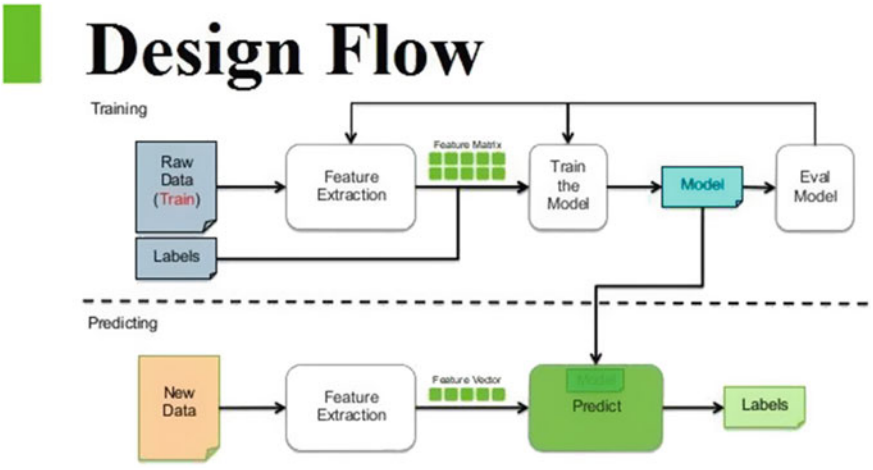


Fig. 3 Design flow

drksujatha23@gmail.com

5 Conclusion

Computers should not be limited to be executing a specific set of instruction to derive a result. It should evolve and discover new ways to solve problems. Using multi-purposed question answer generator, one can save their own time and also save their efforts. They should just give the data to machine and machine automatically generate question and its answer from that data. So, it is very useful to everyone.

6 Future Scope

Currently, we are working on statements and paragraphs. In future, the proposed technique can be improved by we can also work on comprehension, articles, chapters, and also applied on search engine.

References

1. Furnkranz J, Gamberger D, Lavrac N (2012) Machine learning and data mining. In: Foundation of rule learning. Springer (2012)
2. Ikeda M, Ashley K, ArikIturri T-W (2006) An automatic question generator based on Corpora and NLP techniques. Springer, Berlin, Heidelberg
3. Collins M (2002) Machine learning methods in natural language processing. MIT CSAIL
4. Sammons M, Srikumar V (2016) An introduction to machine learning and natural language processing tools. MIAS
5. Donalek C (2011) Supervised and unsupervised learning Ay/Bi 199, Apr 2011

drksujatha23@gmail.com

Building Machine Learning Based Diseases Diagnosis System Considering Various Features of Datasets



Shrwan Ram and Shloak Gupta

Abstract Millions of people worldwide suffer from late diagnosis of diseases. Machine learning algorithms can significantly help in solving healthcare systems that can assist physicians in early diagnosis of diseases. Algorithms in Machine Learning provide the ways to classify data efficiently, at great speed and with high accuracy. Many types of machine learning algorithms are widely adopted and implemented for the early detection of various diseases; these algorithms are like Decision Tree, Naïve Bayes, Support Vector Machine, and Logistic Regression. The results show that there is no particular algorithm available which provides best accuracy in all kind of the healthcare data classification. Most appropriate method can be chosen only after analyzing the nature of the datasets. All the available machine learning techniques are used based on their performances in terms of accuracy and comprehensibility. The datasets considered in this paper are on breast cancer, dermatology, chronic kidney disorder, and biomechanical analysis of orthopedic patients. Data sets from UCI machine learning repository were taken to show applications of Machine Learning on wide variety of Life Sciences data. The four algorithms are implemented with considering various parameters of classification.

Keywords Machine learning · Diseases diagnosis · Supervised learning

1 Introduction

The Machine Learning is all about developing mathematical, computational, and statistical methodologies for finding patterns in and extracting insight from data. Data, in turn, are the concrete manifestations of structures and processes that shape the world. Machine Learning research aims to unlock technologies that can solve

S. Ram (✉) · S. Gupta
M. B. M. Engineering College, Jodhpur, India
e-mail: shrawanbalach@jnvu.edu.in

S. Gupta
e-mail: shloakgupta@gmail.com

© Springer Nature Singapore Pte Ltd. 2019

V. S. Rathore et al. (eds.), *Emerging Trends in Expert Applications and Security*,
Advances in Intelligent Systems and Computing 841,
https://doi.org/10.1007/978-981-13-2285-3_19

drksujatha23@gmail.com

hitherto intractable problems and transform human life in many different areas. Such has already been realized to spectacular effect many times over [1].

Healthcare is rife with rich data and difficult problems; it is therefore fertile ground for machine learning. Indeed, Machine Learning occupies an ever-growing role within health care. A lot of algorithms are developed and deployed for the efficient care of many types of diseases. Millions are suffering due to lack of early discovery of diseases or lack of the knowledge of doctors at various levels. There has always been the demand for intelligent systems having the higher computing power and equipped with sophisticated disease diagnosis models when provided with the various attributes of patient datasets for proper diagnoses [2].

In various cases, patients when diagnosed with a disease it is already too late to start a treatment or in other cases a false judgement by a medical practitioner leading to loss of life. Machine Learning is an option which can analyze a lot of parameters of diagnosis which a doctor can miss in his 10 min meeting with a patient. In this paper, supervised machine learning algorithms are implemented on the healthcare datasets taken from UCI machine learning repository. Datasets are of Breast Cancer, Chronic Kidney Disease and Dermatology. Classification algorithms Naïve Bayes, Decision Tree, Logistic regression, and SVM were used to get results on data mining tool WEKA.

2 Related Work

Many other papers have also been written on diseases diagnosis through machine learning. Like Hao Chan from Yale University who used Deep Neural Network for classification of Melanoma, a kind of skin cancer. He got an accuracy which is equivalent to what a good dermatologist would get [3, p. 6].

In another paper, Kathleen H. Miao and Julia H. Miao from Cornell University used Adaptive Boosting algorithms to correctly diagnose Coronary Heart Disease. They used fourdatasets of different hospitals and got an accuracy ranging from 80 to 90% [4].

Work for early diagnosis of Alzheimer Disease is done by a research group from the University of Sydney where they used Support Vector Machine to get an accuracy of 77% and then they showed that this could be improved using Neural Networks to an accuracy of 83.75% [5, p. 3].

It has been observed that Vocal Chords are the first to be affected by Parkinson Disease. Researchers from two different colleges from Israel showed that Parkinson Disease diagnosis can be automated as Convolution Neural Networks can be used on continuous speech for detection [6].

People from different Universities collaborated in paper which shows that Convolution Neural Networks can be used to classify blood smears as affected or not affected with Malaria. They got an accuracy of 97% in the diagnosis of Malaria through blood smears [7].

drksujatha23@gmail.com

3 Classification Algorithms

Following Supervised Machine Learning algorithms used in this paper.

3.1 Support Vector Machine

Support Vector Machine (SVM) algorithms are very popularly used for the classification task in machine learning; they are based on statistical learning methodology. In SVM, the optimal boundary that separates various classes is known as hyperplane, of two sets in a vector space is obtained independently on the probabilistic distribution of training vectors in the set. The hyperplane is chosen so that the distance from it to the nearest data point on each side is maximized. The vectors close to the hyperplane are called supporting vectors. If the multidimensional space that is formed by plotting of the training data is not linearly separable than this data in its current form does not have a hyperplane. Kernel functions are used in Support Vector Machine (SVM) to solve this problem. These functions analyze the relationship in the data and create complex divisions in space to make categories linearly separable.

3.2 Logistic Regression

Logistic Regression is a standard classification technique based on the probabilistic statistics of the data. It is used to predict a binary response from a binary predictor. Let us assume our hypothesis is given by $h_{\emptyset}(x)$. We will choose

$$h_{\emptyset}(x) = g\left(\emptyset^T x\right) = \frac{1}{1 + e^{-\emptyset^T x}}$$

where is $h_{\emptyset}(x)$ is called the logistic function or the sigmoid function. Assuming all the training examples are generated independently, it is easier to maximize the log-likelihood. Similar to the derivation in case of standard linear regression, we can use any gradient descent algorithm to achieve the optimal points. The updates will be given by $\emptyset := \emptyset - \alpha \Delta_{\alpha} l(\emptyset)$, where $l(\emptyset)$ is the log-likelihood function.

$$g(z) = \frac{1}{1 + e^{-z}}$$

In our use of the logistic regression, we have used L-2 regularization along with tenfold cross validation on the training dataset [8].

drksujatha23@gmail.com

3.3 J48 Decision Tree

In order to classify an item through this supervised learning algorithm, a decision tree is created using the attributes of the training data. Decision tree works for both continuous and categorical attributes. During the training process where the tree is constructed, attributes that differentiate various instances most clearly are identified and are used higher up in the tree. This feature which helps us to classify the instances most clearly is set to have provided highest information gain. Now, if the data instances within its category have the same value for the target variable, then that branch is terminated and it is assigned the target value that we have obtained.

For the other cases, attribute which provides highest information gain is selected. This process is continued until a combination of attributes of the training data gives a particular class or we run out of attributes. In some cases, attributes in a tree can be reused in these trees, the height of the tree can be a limitation. If in case all the attributes are used and still have not reached an unambiguous result from the information, then this branch is assigned a target value equivalent to the majority of the items under the branch (Fig. 1).

3.4 Naïve Bayes

Naïve Bayes classifiers are simple probabilistic classifiers which are based on the concept of Bayes theorem with strong independent assumption. A descriptive term

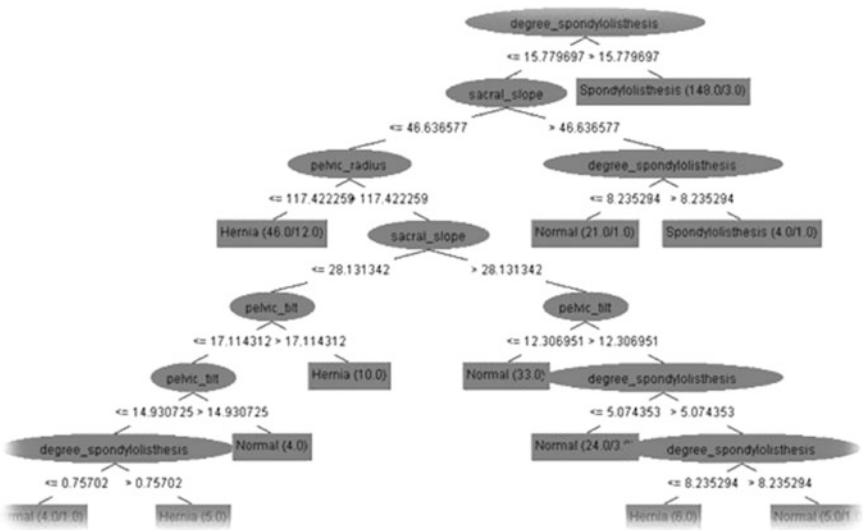


Fig. 1 Trained tree on orthopedic patient dataset

drksujatha23@gmail.com

for this probability model can be a self-determining feature model. Naïve Bayes classifier assumes that the presence of a particular feature of a class is not related to any other feature present in that class. They perform well even if some of the features of the data are related to each other. The main advantage of Naïve Bayes classifier over others is that it requires very less training data to estimate mean and variances of variables needed for classification. Because of the independent assumption of attributes, only the variances of the variables for each label need to be determined during training and not the entire covariance matrix [9].

4 Experimental Setup and Results

All the data used in this paper was taken from UCI machine learning repository and results were obtained on the Machine Learning tool WEKA. The three datasets used are as follows.

4.1 Dermatology

Classification and differential diagnosis of erythemato-squamous diseases is a difficult problem in dermatology. The diseases in this group are pityriasis rosea, lichen planus, psoriasis, chronic dermatitis, seboreic dermatitis, and pityriasis rubra pilaris. These diseases look very similar to each other at the first look with the presence of erythematic and scaling. Detailed analyses may provide us some clinical features like predilection sites or typical localizations which help in identification of particular diseases.

At first, 12 clinical features are used to evaluate the patients. These features may provide the insight needed to classify the diseases but generally biopsy is done to get more accurate diagnosis. In biopsy 22 histopathological features are collected from every patient. Another problem in diagnosis of erythemato-squamous diseases is that a disease may show the histopathological features of another disease at the initial stages and may change to its particular characteristics in further stages [10].

In this paper, we have used machine learning algorithms mentioned in this paper on the data set taken from UCI machine learning repository in the first task and analyzed only the clinical features and then in the next task, analyzed for all the attributes. The results are shown through this graph in Fig. 2.

The blue bars show the accuracy of first task when only clinical attributes were used and the red bars show accuracy when all attributes were used. There is approximately 10% increase in accuracy when histopathological features were added. Naïve Bayes and Logistic Regression show the highest accuracy of 97% on

drksujatha23@gmail.com

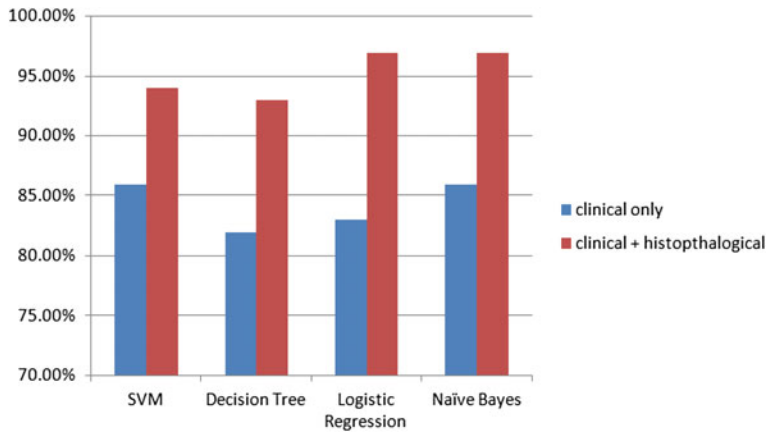


Fig. 2 Graph showing accuracy on Dermatology dataset

all attributes. Confusion matrix obtained from Logistic Regression showed that there is need for more data so to differentiate seboreic dermatitis and pityriasis rosea.

4.2 Breast Cancer

The second leading cause of death among women is breast cancer. It is now the most common cancer to be found in most cities in India, and second most common in the rural areas. Early and correct diagnosis of cancer will help in saving life. Applying machine learning algorithms on breast cancer dataset taken from UCI machine learning repository to classify if the tumor is benign or malignant.

This dataset consists of 212 malignant and 357 benign instances, where each one represents FNA test measurements for one diagnosis case. For this dataset, each instance has 32 attributes, where the first attribute is the identification number and second attribute in the status of cancer (benign/malignant) [11, p. 7]. The remaining 30 features are ten real-valued features with their mean, standard error and mean of three largest worst values for every nucleus. These ten values are calculated from a digitized image of a fine needle aspirate (FNA) of breast tumor, extracting the features from the cell nuclei in the image [12].

There are a lot of correlated features in this dataset like radius perimeter and area. Removing these correlated attributes and unique Ids before classifying on remaining 25 attributes (Fig. 3).

These classifiers when coupled with principal component analysis (feature selection) gave the best accuracy of 97%. Accuracy this high proves that machine learning is a great prospect for early diagnoses of breast cancer.

drksujatha23@gmail.com

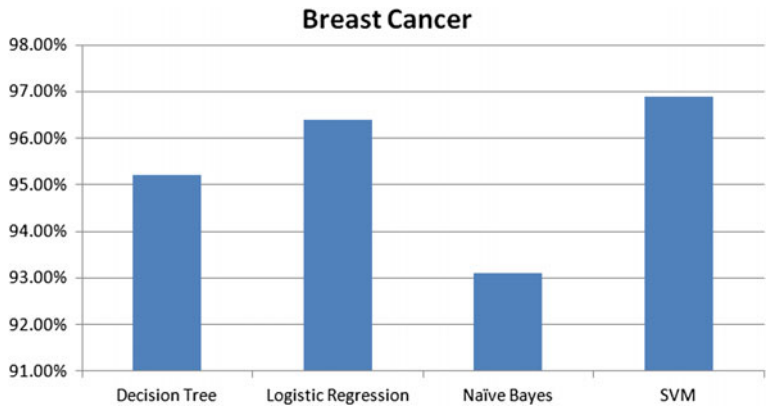


Fig. 3 Graph showing accuracy on Breast cancer dataset

4.3 Chronic Kidney Disease

Chronic Kidney Disease is a worldwide health problem with an increase in its prevalence. In the past decade, numbers of patients that have CKD have almost multiplied by a factor of 3. Of the total population of this world, 5–11% of the people may have some form of CKD. CKD also promotes dyslipidemia and hypertension. There are very little symptoms in early stages of CKD which makes early detection very difficult. Studies have come up which have shown how poorly local medicine practitioners fair out in diagnosing CKD and the results are depressing. More than 50% of the CKD cases go undiagnosed in primary stages of it. Hence, an accurate and convenient way is needed which may help physicians in early diagnosis of CKD [13]. In this paper, an automated Machine Learning solution is developed to detect CKD (Fig. 4).

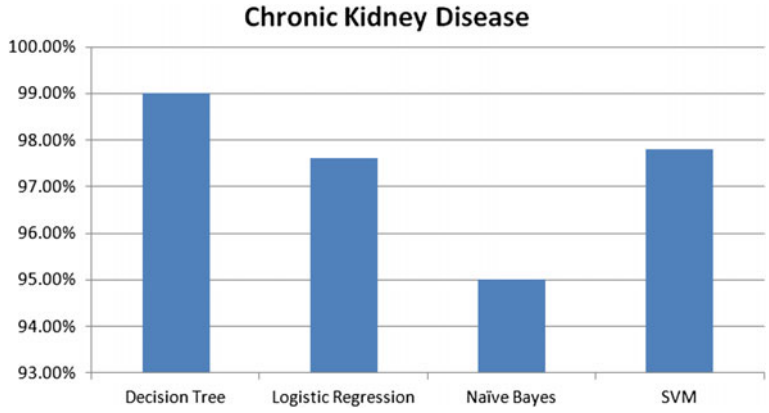


Fig. 4 Graph showing accuracy on CKD dataset

drksujatha23@gmail.com

Accuracy of 99% was achieved using J48 on CKD dataset from UCI machine learning repository. This data set had features like age, cell counts, sugar level, hemoglobin, potassium, sodium, urea levels, etc. This shows that intelligent machine can now be used for diagnosis of CKD.

5 Conclusion

- On dataset for diagnoses of erythemato-squamous diseases logistic regression and Naïve Bayes performed best with 97% accuracy. Histopathological features are important as including them increase in classification accuracy of 10% is achieved.
- After feature selection using Principal Component Analysis, SVM gave an accuracy of 97%. On breast cancer dataset.
- 99% accuracy was achieved when J48 decision tree algorithm was applied on chronic kidney diseases dataset.
- High accuracy on varied healthcare datasets used in this paper provides enough motivation to build intelligent systems for health care. Early diagnosis is possible and loss of life due to human inefficiency will decrease. Increase in computing power of machines now days will allow the use of more complex machine algorithms and increase in processing speed even if large attributes are present in datasets.

6 Future Scope

In this paper, we have obtained high accuracy on these three datasets but higher accuracy can be achievable if more data is available. More parameters like age, sex, locality or others can be added to the data sets so accuracy can be improved. Standardized attributes can be set across clinics where patients are tested to have uniform data. This paper had numerical data which in some cases like breast cancer are taken through feature extraction from images. In future classification algorithms like Convolution Neural networks can be used to automate this extraction. In all, there is a bright future for Machine Learning aiding doctors in improving patient quality of life.

drksujatha23@gmail.com

References

1. Fatima M, Pasha M (2017) Survey of machine learning algorithms for medical diagnosis. J Intell Learn Syst Appl
2. Jain A (2015) Machine learning techniques for medical diagnosis: a review. In: Conference on science, technology and management
3. Chang H (2017) Skin cancer reorganization and classification with deep neural network, 6
4. Miao KH, Miao JH (2016) Diagnosing coronary heart disease using ensemble machine learning. Int J Adv Comput Sci Appl (IJACSA)
5. Liu S, Liu S (2014) Early diagnosis of Alzheimer's disease with deep learning. IEEE, 3
6. Frid A, Kantor A (2016) Diagnosis of Parkinson's disease from continuous speech using deep convolutional networks without manual selection of features. In: ICSEE international conference on the science of electrical engineering
7. Liang Z, Powell A (2016) CNN-based image analysis for malaria diagnosis. In: International conference on bioinformatics and biomedicine. IEEE
8. Khanna D, Sahu R (2015) Comparative study of classification techniques (SVM, logistic regression and neural networks) to predict the prevalence of heart disease. Int J Mach Learn Comput 5(5)
9. Vijayarani S, Dhayanand S (2015) Liver disease prediction using SVM and Naive Bayes algorithms. Int J Sci Eng Technol Res (IJSETR)
10. Priyadharshini J (2015) A classification via clustering approach for enhancing the prediction. Int J Sci Res Dev 3(06)
11. Zafiropoulos E (2006) A support vector machine approach to breast cancer diagnosis and prognosis
12. Salama A (2012) Breast cancer diagnosis on three different datasets. Int J Comput Inf Technol
13. Salekin A, Stankovic J (2016) Detection of chronic kidney disease and selecting

drksujatha23@gmail.com

Enhancing Data Security in Cloud Using Split Algorithm, Caesar Cipher, and Vigenere Cipher, Homomorphism Encryption Scheme



Abhishek Singh and Shilpi Sharma

Abstract Cloud computing declared among the fastest growing technology. The amount of users using this technology has exploded. Therefore, cloud users expect their data to be safe on the cloud. We can secure the data from being accessed illegally by encrypting it with a key with various methods likewise Advanced and Data encryption standard. However, the issue with these encryption methods is that they are very complex and have a lengthy computing time due to finite key space. To overcome these problems, Cloud Service Providers (CSP) which is becoming a great area to have researched for ensuring the security of data with minimum processing time. This research paper emphasizes various secured encryption techniques which have efficient computing time. This paper focus on the complexity, efficiency, and the security of the algorithm used.

Keywords Cloud computing · Encryption · Data security · Split algorithm Caesar cipher · Vigenere cipher

1 Introduction

Cloud computing declared among the important technology is being used in the world, which is allowing corporate organizations to have an access to different apps, data stores that too without accessing the user personal details/files uploaded [1] Considering the capacity, sturdiness, and the reliability of the cloud, we cannot ignore the various threats to the user's data that is uploaded on the cloud. Untrusted

A. Singh (✉) · S. Sharma
Amity School of Engineering and Technology, Amity University Uttar Pradesh, Noida, Uttar Pradesh, India
e-mail: abhishekbanshal27@gmail.com

S. Sharma
e-mail: ssharma22@amity.edu

© Springer Nature Singapore Pte Ltd. 2019

V. S. Rathore et al. (eds.), *Emerging Trends in Expert Applications and Security*, Advances in Intelligent Systems and Computing 841,
https://doi.org/10.1007/978-981-13-2285-3_20

drksujatha23@gmail.com

cloud servers compromise the security of the cloud and allow access to unauthorized users. File entrance mechanism is another challenging issue in cloud storage system [2]. It produces redundant copies of similar files and data. Attacks from unauthorized users are hard to stop in cloud computing storage.

In this paper, we are proposing the concept of storing data on multiple cloud servers using the technique of encryption instead of storing a whole file on one system. This system will segregate the files into multiple chunks of data and encrypt it using the various encryption techniques discussed on the paper and then storing it on multiple clouds. The data important for decryption and rearrangement of that file gets stored on a management server based on metadata for effective and correct retrieval of the original file.

The idea behind this design is to divide the logic applied into multiple parts and parts get distributed to particular cloud. The customer encrypted the information with the help of open key and transfer the scrambled information on to the cloud. The cloud then tries to figure out the encoded information to get a scrambled outcome that allows the client who transferred the record can decipher and nobody else can. The client or a trustful cloud deals with the encryption key and plays out the different operations while the gigantic calculation on scrambled information is done by an untrusted cloud.

2 Related Work

Vijay G. R. also, A. Rama Mohan Reddy have proposed a proficient security display in distributed computing condition with the assistance of delicate processing systems. Here, a solid security in distributed computing is made do with the assistance of the administration framework to guarantee the information security. It is additionally keeping up the exchange table that contains the information.

Jeffrey Holmes and Nandita Sen Gupta have discussed enhancing the security system for cloud computing using cryptography [4]. They proposed an efficient cryptography system named as HVCCE. The framework will keep the foundation of the cloud in essential places, for example, the customer's area in the server and the system. The encryption contains 3 phases.

Singh and Supriya [5] have redesigned the Vigenere encryption calculation. Their mixture usage with base 64 and AES has proposed another approach for utilizing the encryption calculations with various sort of calculation, for example, substitution figure, symmetric calculations and so forth.

Sindhuja and Devi [6] in their paper examined on symmetric key encryption strategy utilizing nonspecific calculation key have proposed a hereditary calculation that is based onto the symmetric key cryptosystem that is used for encryption and decoding.

drksujatha23@gmail.com

3 Methodology

Following algorithm included also perform encryption and decryption with the help of the given modules:

- Caesar cipher encryption: The plaintext data is loaded and the encryption converts it into a cipher which is one of the levels of encryption being used.
- Vigenere cipher: This step implements the conventional Vigenere through which we can generate the next level of ciphertext but it is generated all in upper case as the conventional algorithm is not case-sensitive.
- Vigenere decipher: Reverse process is used for the decryption.
- De-Caesar cipher: Again in order is used to obtain real plaintext, we perform Caesar cipher in reverse manner and obtain the plaintext.

3.1 Split Algorithm

Split algorithm is a technique used for security of information over a network. It involves encryption of data, splitting the data into smaller units and then distributing it to multiple storage locations and the further encrypting the new information on its new location [7]. It permits numerous offers; no single circle has every one of the information. Virtualization and encryption give various COI's on a circle give more proficient utilization of capacity. Information encoded with a solitary key is constrained. More secure Redundancy: RAID 5 calculation gives extra data to assailants [8]. It has more prominent many-sided quality in the SAN and arrangement. Be that as it may, this repetition calculation expends more storing facility.

3.2 Vigenere Cipher

The Vigenère cipher is the strategy for encryption of content by utilizing Caesar cipher in light of the content of a catchphrase [9].

3.3 Caesar Cipher

The Caesar Cipher is declared among the least complex figure. It is a kind of substitution method of ciphertext onto which each text gets moved into specific number of spots bottom to text in order [10]. The upside of both Vigenere and Caesar cipher is that gives triple encryption to the records. It is one of the most effortless technique to use in cryptography and can give security to the information.

drksujatha23@gmail.com

Utilization of just a short key in the whole procedure. A noteworthy disadvantage is that these figures are less proficient. It can just give the least security to the data. Recurrence of the example gives a gigantic piece of information in deciphering the whole text [11].

3.4 Homomorphism Encryption Scheme

Homomorphic encryption is a type of encrypted data that permits calculation of figure writings, producing an encoded result which, when decrypted, matches the estimation of operations performed on the plaintext [12]. The benefit of this strategy is that it successfully prompts the security of information exchange and the information storage of cloud framework. It is utilized as a part of managing account exchanges, voting frameworks, distributed computing applications, and private data recovery. The disservice is at exhibit, completely homomorphic encryption conspire has a high calculation issue needs encourage study [13].

4 Proposed Model

4.1 File Encryption Technique Module

The word encryption does not entertain interference but instead rejects a message to some interceptor. This encryption area utilizes the random structure by which an encryption key gets created with the help of calculations. It is almost difficult to try decoding information without the use of key and at this similar time for the created encoding some abilities must be required. With the approval, decrypted message can be retrieved, however not to unapproved clients. In our proposed system, we utilize mix of AES calculation and some parts of SHA-1 calculation for the encoding the part of file.

4.2 Module and Decryption Technique

Decrypting gets utilized to help to clarify the steps for the procedure of decoding the stored information which is physically stored or utilizing the best possible keys.

Stored information gets decrypted to make it capable of giving troublesome another person for taking the clients' information illicitly. There are some organizations who try to encode the given information for some general assurance of information of the organization. This information must be distinguishable. This might required decoding. In our system that key which is used for decoding cannot be accessed, some exceptional method is required to perform decoding that information for breaking the un-encoding and made the information comprehensible.

drksujatha23@gmail.com

4.3 Split and Club

In this proposed system, we try to deviate the file into some different parts and then tries to encrypt and tries to get it stored onto the multiple cloud. Here our plan makes use of as many cloud as possible at the same time tries to avoid the risks of many unauthorized activity, manipulation of knowledge access done illegally and some method meddling. Our proposed design will try to target the confidentiality of information. Our idea behind this is that the logic should be deviated into components and these components must be deviated onto multiple cloud.

4.4 AES

AES depends on the standard of 'substitution-change organize', in which blend of every substitution and stage is snappy in each bundle and equipment [8].

It is unique in relation to DES as AES cannot utilize a structure described by the fiestel. AES might be described as the variant of Rijindael that contains a square size of almost 128, and key size of around 128, 192, 256. Against this,

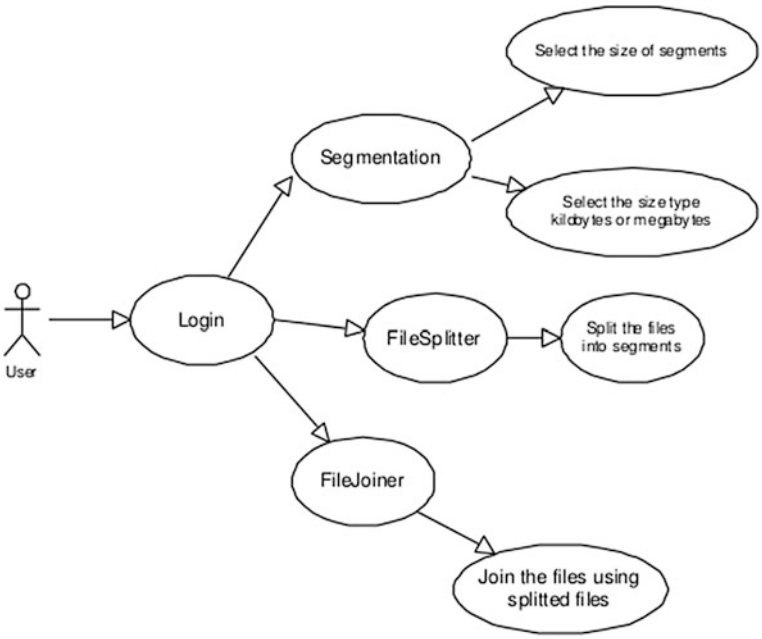


Fig. 1 Use case diagram

drksujatha23@gmail.com

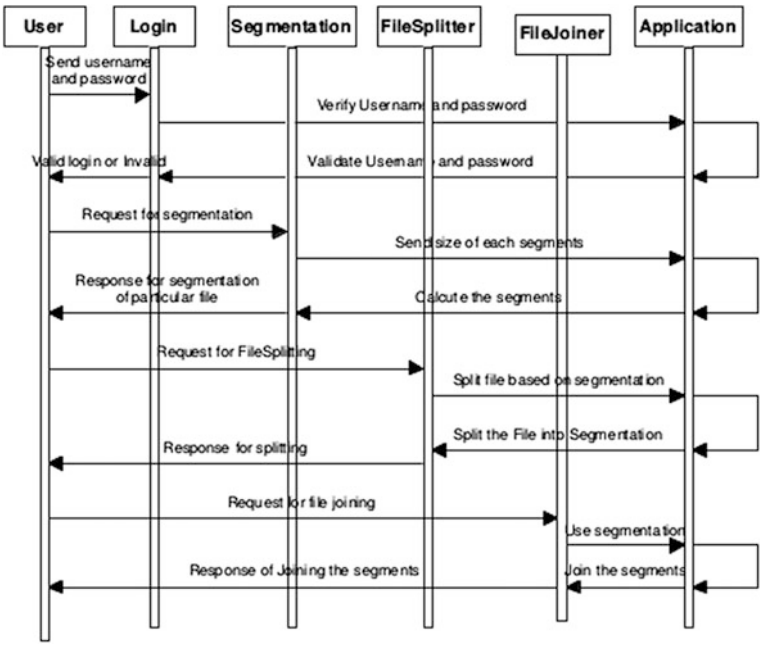


Fig. 2 Sequence diagram

determination by Rijindael all by itself is particular with a key size which will be anyhow different of 32 bits, each of at least 128 and a large portion 256 bits. All by itself is particular with the piece and key sizes which will be any different of 32 bits, each with at least 128 and a large portion of 256 bits.

AES tends to work on a 4×40 column noteworthy lattice request of bytes, named the position, then also some adaptations by Rijndael tends to have some greater size of square (Figs. 1 and 2).

5 Result

5.1 Result Definition

The outcome of this paper is to provide security, this system needs to analyze some famous risks, and it needs to specify measures like encoding for providing security, cost. The copies of software which are kept as a backup file and procedure needs to be available for recovery restart whenever start.

drksujatha23@gmail.com

5.2 Security Against Unauthorized Access

5.2.1. **Administer Password**
The password has to be used because it will be providing security for the client because it will stop any malicious activity.
5.2.2. **Validation and Checks**: The user and developer will check user related checks and some validation that will be from the user.
5.2.3. **Authorized Key**: In order to get login password from the user will be checked.

5.3 Secured for Data Loss

5.3.1. **Data Backup System**: In our proposed system, a new enhanced system will be used that will be capable of providing enough backup.
5.3.2. **Offline Service**: In this proposed system, data can be stored offline.
5.3.3. **Different Database Backup**: In our proposed system, different databases will be used to provide backup.

Table 1 shows the comparison done between the different algorithms like split algorithm, Vigenere cipher, Caesar cipher, and Craig Gentry based upon the parameters platform, security, privacy, key used, and efficiency.

Vigenere Cipher

See Fig. 3.

Output:

Ciphertext: azhgfnikqvtkh. Original/Decrypted Text: abnoynkqyotmn.

Table 1 Comparision of algorithms

Parameters	Split algorithm	Caesar and Vigenere cipher	Craig Gentry
Platform	Cloud	cloud	cloud
Type	Encryption with a single key	Triple	Fully
Security [10]	Cloud server	Cloud server	Cloud server
Privacy of data	Storage and communication	Storage and communication	Storage
Keys used	Single key is used for coding and decoding of ciphertext	Different keys are used	Different key
Efficiency [11]	Highly efficient	Less efficient	Very less efficient

drksujatha23@gmail.com

Caesar Cipher

See Fig. 4.

Output:

Text: attackatonce. Shift: 4. Cipher: exxegoexsrgi.

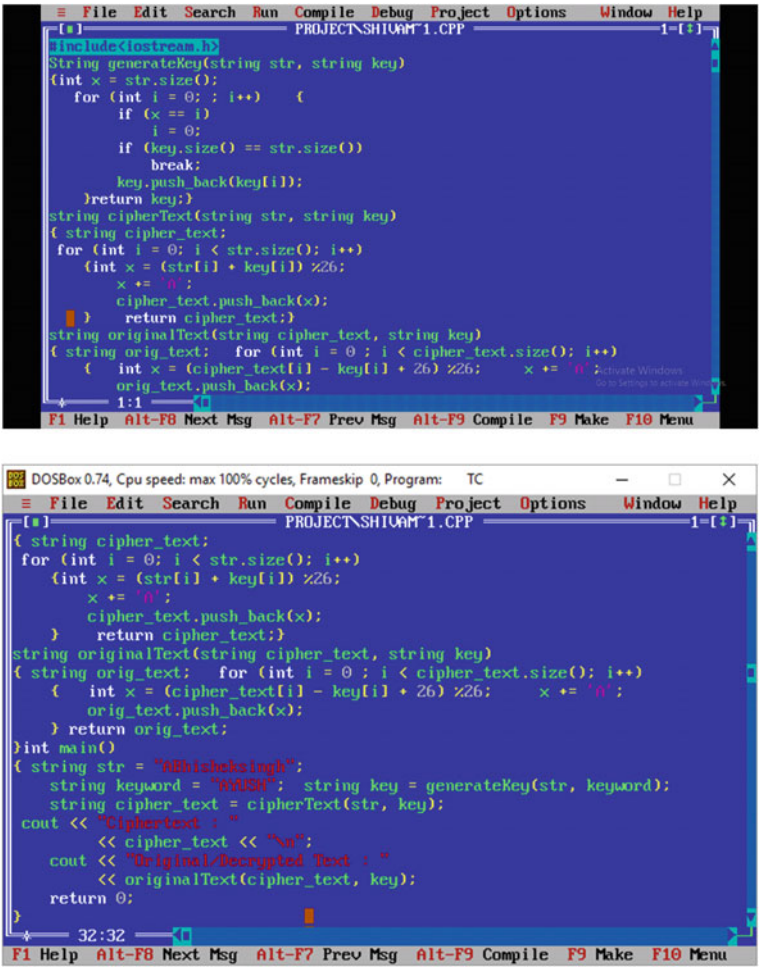


Fig. 3 Implementation of Vigenere cipher

drksujatha23@gmail.com

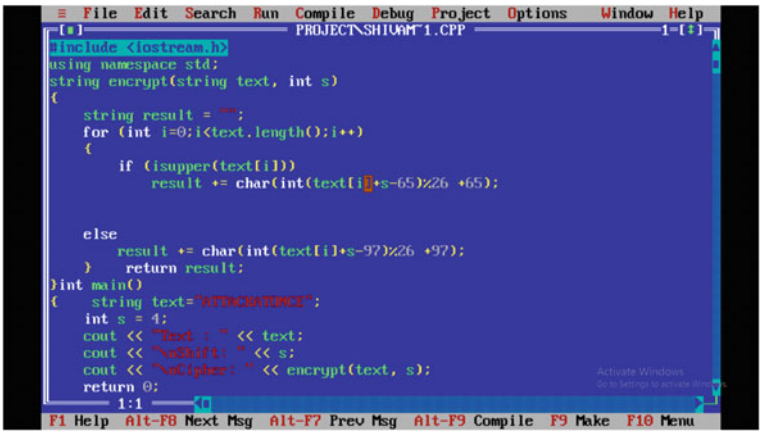


Fig. 4 Implementation of Caesar cipher

6 Conclusion

In this, we have studied an arrangement of cloud information storage, clients store their information in the cloud, so that there is no need to store them on their neighborhood frameworks. Accordingly, the security, honesty, and accessibility of information documents on storage distributed cloud servers are normal. To consent to this, the framework and arrangements of included components in the information stockpiling in the cloud framework ought to be checked. About the customer, we should utilize an encryption from the client like AES encryption, its high security and protection have been tried in many testing. AES has been analyzed and assessed by the NIST and its security has been affirmed by this Institute. This encryption is utilized to scramble exceedingly delicate data in the United States of America. We can likewise utilize new strategies like hereditary or dynamic calculation encryption calculation which can build the encryption drastically.

The following component is the server; our data has been put away on the server and we have the storage room remotely as a client. Accordingly, the accessibility of the data and its recovery is extremely noteworthy and ought to give all the ways to secure the server side. To agree to this, we think about between a few arrangements by suppliers for security known in the field of information storage administrations. The examination we canned be obviously observed that with a specific end goal to the classification of data, a few suppliers utilize the encryption control instrument, for example, symmetric encryption [14]. About the security of our server specialist co-ops in this field to extend and to build the security instruments on their servers, on the grounds that the clients of cloud innovation will go to the side of those suppliers that their administrations have enough security, along these lines server security will be essential and suppliers would success be able to in this innovation with high server security and responsibility to the clients. The third component is

drksujatha23@gmail.com

that its security is essential in the capacity and transmission of information is the association channel between cloud specialist organizations and client. For this situation, we can allude to the conventions and setting up more secure transmission channels that they create by utilizing new and techniques in the area [15]. Numerous organizations have protected and secure capacity issues that are explained by actualizing the cloud-based capacity. The proposed framework will give the clients an entire security of its information.

References

1. Vijay GR, Rama Mohan Reddy A (2012) An efficient security model in cloud computing based on soft computing techniques. Int J Comput Appl 60(14)
2. Xiao Z, Xiao Y (2013) Security and privacy in cloud computing. IEEE Commun Surv Tutor 15(2):843–859
3. Singhal M et al (2013) Collaboration in multicloud computing environments: Framework and security issues. Computer 46(2):76–84
4. Sengupta N, Holmes J (2013) Designing of cryptography based security system for cloud computing. In: 2013 International conference on cloud and ubiquitous computing and emerging technologies (CUBE). IEEE
5. Singh G, Supriya (2013) Modified vigenere encryption algorithm and its hybrid implementation with Base64 and AES. In: 2013 2nd International conference on advanced computing, networking and security (ADCONS). IEEE
6. Sindhuja K, Devi PS (2014) A symmetric key encryption technique using genetic algorithm. Int J Comput Sci Inf Technol 5(1):414–416
7. Almarimi A et al (2012) A new approach for data encryption using genetic algorithms. Adv Intell Syst Comput 167:783–791
8. Shah Kruti R, Gambhava B (2012) New approach of data encryption standard algorithm. Strings 1
9. Thirer N (2013) A pipelined FPGA implementation of an encryption algorithm based on genetic algorithm. In: SPIE defense, security, and sensing. International Society for Optics and Photonics
10. Siddiqa A et al (2016) A survey of big data management: taxonomy and state-of-the-art. J Netw Comput Appl 71:151–166
11. Bono SC et al (2014) Systems and methods for secure workgroup management and communication. U.S. Patent No. 8,898,464. 25 Nov 2014
12. Bruen AA, Forcinito MA (2011) Cryptography, information theory, and error-correction: a handbook for the 21st century, vol 68. Wiley
13. Soni G, Gupta U, Singh N (2014) Analysis of modified substitution encryption techniques, pp 643–647
14. Micciancio D (2010) A first glimpse of cryptography's Holy Grail. Commun ACM 53(3):96–97
15. Yang K, Jia X (2012) Attributed-based access control for multi-authority systems in cloud storage. In: 2012 IEEE 32nd international conference on distributed computing systems (ICDCS). IEEE

drksujatha23@gmail.com

k-dLst Tree: k-d Tree with Linked List to Handle Duplicate Keys



Meenakshi and Sumeet Gill

Abstract Spatial data can be indexed in many ways like indexing of points, lines, and polygons. And, there are many indexing structures like R-tree, k-d tree, grids, and their variants, which are used for indexing geospatial data. The fundamental type of the k-d structure is used to index k-dimensional data. Every interior node of the k-d structure holds a data coordinate and represents a rectangular area. Root of the k-d tree structure represents the whole area of interest. The k-d tree is a main memory structure. Though the main memory methods are not designed to handle very large datasets, these data structures show many interesting features for handling spatial data. The spatial datasets might have several records for the same spatial location. In this paper, we are proposing the novel indexing structure k-dLst tree to index the spatial records with duplicate keys. The proposed indexing tree is based on k-d tree indexing structure.

1 Introduction

For accessing and manipulating the spatial data resourcefully, which is the basic requirement of some geodata-based applications; a system tremendously requires an indexing technique to facilitate it in access and retrieval of data with a speed based on their geospatial position. These systems require spatial search on multidimensional spaces. The conventional techniques used for indexing are not suitable for storing objects whose size is not zero and exist in n-dimensional regions [1]. To support speedy retrieval of data from big-sized datasets is the main expectation from indexing data structures. The indexing is done by keeping in mind the preservation of spatial relations like covers, overlaps, nearest neighbor, etc., while indexing objects. Through the years, many tree-based indexing structures are

Meenakshi (✉) · S. Gill
Department of Mathematics, M. D. University, Rohtak, India
e-mail: mshthebest@gmail.com

S. Gill
e-mail: drsumeetgill@gmail.com

© Springer Nature Singapore Pte Ltd. 2019

V. S. Rathore et al. (eds.), *Emerging Trends in Expert Applications and Security*,
Advances in Intelligent Systems and Computing 841,
https://doi.org/10.1007/978-981-13-2285-3_21

drksujatha23@gmail.com

designed and proposed to index spatial entities like R-tree, Quadtree, Grids, etc. These structures offer well-organized indexing techniques to index n-dimensional objects while keeping the spatial aspects in mind. Many other indexing structures are also found in the literature which are variants of basic spatial structures like the R+-tree, R*-tree, etc. In our research work, the k-d tree is used as a basic structure to index geospatial data. In this paper, we are proposing a novel indexing structure k-dLst tree.

2 k-d Tree

Bentley [2] proposed an indexing structure k-d tree in 1975. It is based on generalization of binary search tree for supporting data related to multiple dimensions. It considers a set of k-dimensional keys that need to be stored efficiently. Every node of the kd-tree structure has one of the keys and one discriminating value related to it. Discriminating values ranges between integers from 0 to $k - 1$. To begin with, the node at the root corresponds to the entire space under consideration. Let k_root be the key at the root and dc be its discriminating value. Then, k_root with respect to dc will divide the entire space into two areas such that all the keys ky less than k_root for the same discriminating value will go into the left subtree, and all the keys ky either greater than or equal to k_root for the same discriminating value will be inserted into the right subtree. The same technique to divide the whole space is followed for all subtrees recursively, until the empty trees are reached. The discriminating value at each node is updated by one in a circular way from 0 to $k - 1$. It will be chosen by alternating the coordinates of each level, starting with 0: first using dimension 0 for the root, then dimension 1 for the next level, then dimension 2 for one level further, …, then using the dimension $k - 1$ at the kth level, and then restarting at dimension 0 again, etc. [3].

3 Spatial Indexing for Duplicate Keys

The primary function of the spatial indexing structures is to provide proficient and speedy retrieval of data items on the basis of their spatial features. For instance, a point query shows the results about objects at particular location, a range query about the objects which fall in given range of geospatial coordinates and a query to inspect about nearest neighbor looks for the object(s) which is nearest to a particular object. The main characteristics of spatial datasets include their large size and irregular distribution. If we try to access any object without indexing the data, it will require each entity in the database or dataset to be inspected to conclude whether it fulfills the condition. A complete table or dataset scan in a database or dataset is required. Due to the reason that spatial datasets are frequently very large, such kind of full scans is unacceptable practically, especially in the case of interactive

drksujatha23@gmail.com

applications. So, an indexing method to index spatial data is definitely required to locate the needed objects proficiently without traversing each and every entity.

More often, in spatial databases, we do have multiple entries for the same spatial location. Many researchers either do not consider such kind of data or have remained silent about handling of multiple entries related to the same spatial key. Brown [4] handles the duplicated keys in the same way as the keys which are less than the compares key are handled, i.e., the algorithm considers both *less than* (<) and *equal to* (=) relations with respect to the comparison of keys in the same way. In [5], duplicated tuples are first removed and then, the remaining data is considered for further processing. https://www.cs.cmu.edu/~ckingsf/bioinfo-lectures/kdtrees.pdf [6] considers the duplicate keys as an error and no further processing is done for such data. https://www.cs.umd.edu/class/spring2002/cmsc420-0401/pbasic.pdf [7] is silent about handling of duplicate keys. In [2], when a duplicate key is found, the address of already existing node holding the data for the same key is returned back.

There are different ways to handle such records with duplicate keys according to requirements of different applications. Some of these are:

1. Simply ignore the records with duplicate keys.
2. Insert the first record and ignore the others with same key.
3. Insert the first record and update it with new data for the same key.
4. Save all records related to the duplicate keys.

The disadvantage of the first three approaches is the loss of data and the result of queries will not show genuine results. So, we need to index all records to show correct results of queries.

4 Proposed Work: k-dLst Tree

k-dLst tree is based on k-d tree. The point which is not in structure is inserted by going down the tree structure until the end of the tree, i.e., a leaf node is founded. Starting at the root, the coordinate values of every internal node are compared with the respective coordinate values of the new point to be inserted and the right path is selected based on the output of the comparison. The process is repeated until a leaf node is arrived at.

If the data records of a dataset are indexed using k-dLst tree, then there will be a *kdLstNode* for every composite key of k attributes and it will point to *dataNode* containing remaining nonspatial data attributes related to the key. If any key has duplicate data records, then it will be added as a new *dataNode* using linked list. Additionally, *kdLstNode* contains k pointers that are either initialized to *NULL* or address to the other *kdLstNode* in k-dLst tree structure. Every *kdLstNode* might contain a discriminator as a field which holds an integer value between 0 and $k - 1$ in an inclusive way. Figure 1 shows the structure of k-dLst tree for dataset records in 2-d space stored as *kdLstNodes* in a 2-dLst tree, which can be generalized for k-d space to create k-dLst tree. A rectangular space is also represented by the leaf node

drksujatha23@gmail.com

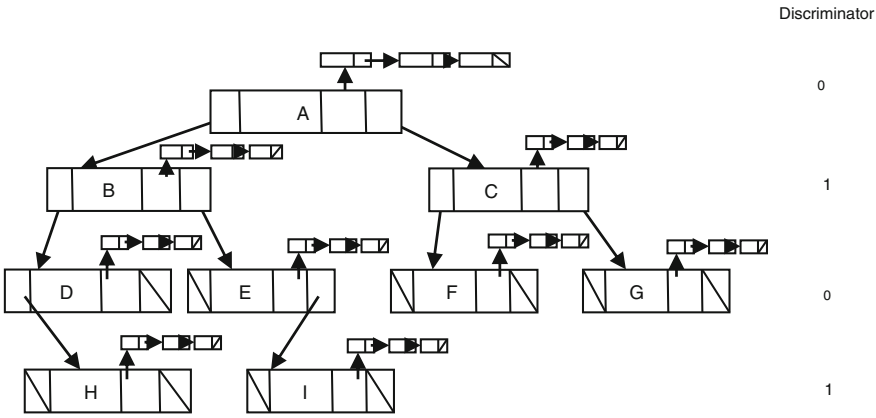


Fig. 1 Structure of k-dLst tree for 2-d keys (It can be generalized for k-d keys)

which is further separated into two spaces by the newly inserted point. When a new node is inserted, it results in one more node which is internal.

5 Explanation

Every node on any particular level of kdLst tree has the same discriminator. The root node will have discriminator 0. It will be 1 for two sons at the next level, and will be incremented for every next level until it reaches up to $k - 1$ on kth level. Then it again starts with 0 for $k + 1$th level and the cycle repeats till the end of kdLst tree.

So, **nextDiscriminator** (*level_i*) = (*level_i* + 1) mod *k*.

Notations:

K	*Keys of kdLstNode*
N	*kdLstNode of k-dLst tree structure*
$\boldsymbol{K_0(N), K_1(N) \ldots K_{k-1}(N)}$	*k keys of kdLstNode N*
***lSon*(*N*)**	*Left branch of kdLstNode N*
***rSon*(*N*)**	*Right branch of kdLstNode N*
***dc*(*N*)**	*Discriminator of kdLstNode N*

Now, whether to insert new *kdLstNode* as left son or right son depends on the result of comparison of keys. Let *dc* be the discriminator for *kdLstNode N*. If $K_{dc}(N)$!= $K_{dc}(Q)$, then the successor *kdLstNode Q* will be inserted either on left or right side of *N*, i.e., *either lSon*(*N*) *or rSon*(*N*). If $K_{dc}(N) < K_{dc}(Q)$, then *Q* will be inserted on right side of *N*, i.e., *rSon*(*N*) else on left side of *N*, i.e., *lSon*(*N*). But, if $K_{dc}(N) = K_{dc}(Q)$, then the remaining part of the keys will be compared. A superkey

drksujatha23@gmail.com

SK of *kdLstNode N* is defined by cyclical concatenation of all keys starting with $K_{dc}(N)$ as

$$SK_{dc}(N) = K_{dc}(N)K_{dc+1}(N)\ldots K_{dck-1}(N)K_{dc0}(N)\ldots K_{dc-1}(N)$$

Now, if $SK_{dc}(Q) < SK_{dc}(N)$, then Q will be added as left son of N else Q will be added as right son of N.

Now, in case of duplicate keys, i.e., if $SK_{dc}(Q) = SK_{dc}(N)$, then address of new *dataNode* will be pointed by the same *kdLstNode* and other *dataNodes* with same key already inserted will be added as linked list next to new *dataNode*. In this way where the keys are same the *dataNodes* create a linked list in such a way that the latest node will be inserted as first node, i.e., head of the *dataNode* list.

5.1 Creation of k-dLst Tree Structure

The algorithm to create and insert a node in k-dLst tree is described here. A new dataset record *lstDataPoint ldp*, node of k-dLst tree/subtree *kdLstNode kln* and *discriminator dc* are passed to INSERT algorithm. If there is no node in k-dLst tree with equal keys, it will insert the *lstDataPoint ldp* in proper *kdLstnode kln* at proper position; otherwise, the *lstDataPoint ldp* will be appended to the list of *kdLstNode kln* with equal keys. **Algorithm 01 k-dLst_INSERT** explains how a node is inserted in k-dLst tree.

Algorithm 01:

k-dLst_INSERT (lstDataPoint ldp, kdLstNode kln, discriminator dc)
If kln IS NULL
Create a new kdLstNode kln
Insert ldp in kln and update the pointers
return
else
If ldp_KEY IS EQUAL TO kln_KEY
Insert ldp as a starting node in the Linked List of kdLstNode kln and return
Else if ldp_KEY < kln_KEY then
Set kln to the left pointer in kln
Update dc
Call k-dLst_INSERT(ldp, kln, dc)
else
Set kln to the right pointer in kln
Update dc
Call k-dLst_INSERT(ldp, kln, dc)
End if
Return kln
END

drksujatha23@gmail.com

5.2 Searching in k-dLst Tree Structure

The algorithm to search for some data at particular spatial location in k-dLst tree structure is described here. The spatial data coordinates *sdp* of *spatialDataPoint* type, *kln* of *kdLstNode* type, and *dc* of *discriminator* type are passed to SEARCH algorithm. In starting, s*dp* holds the geospatial coordinates about which the query is asked, *kln* holds the root of k-dLst tree, and *dc* holds the value of discriminator at particular level (holds 0 for the root node). As the search is continued, further values for *kln* are updated with *lSon*(*kln*) or *rSon*(*kln*). The value of discriminator *dc* is also updated accordingly for every level. **Algorithm 02 k-dLst_SEARCH** explains how the objects at any particular spatial location can be searched using k-dLst tree.

Algorithm 02:

k-dLst_SEARCH (spatialDataPoint sdp, kdLstNode kln, discriminator dc)
[Check whether k-dLst tree is NULL]
If kln IS NULL
Return NIL
else
if sdp_KEY IS EQUAL TO kln_KEY
Traverse and show the full Linked List of dataNodes in kdLstNode and return
Else if sdp_KEY < kln_KEY then
Set kln to the left pointer in kln, i.e., lSon(kln)
Update dc
Call k-dLst_SEARCH (sdp, kln, dc)
else
Set kln to the right pointer in kln, i.e., rSon(kln)
Update dc
Call k-dLst_SEARCH(sdp, kln, dc)
End if
Return NIL

6 Experimental Analysis

The dataset we have used for experimental analysis of k-dLst tree is an artificial dataset created by using a Linux script.

It contains the geospatial 2-d coordinates containing latitude, longitude, and object_id of objects at a particular geospatial location. The dataset holds the records which have same latitude and longitude values, i.e., duplicate composite keys. We see such kind of data frequently in real life situations (Table 1).

drksujatha23@gmail.com

Table 1 Location dataset of objects

Latitude	Longitude	Object_Id
10	90	Obj_Id 1
25	23	Obj_Id 2
26	56	Obj_Id 3
20	45	Obj_Id 4
34	70	Obj_Id 5
10	90	Obj_Id 19
25	23	Obj_Id 20
26	56	Obj_Id 21
20	45	Obj_Id 22
34	70	Obj_Id 23
56	32	Obj_Id 24
10	90	Obj_Id 66
25	23	Obj_Id 67

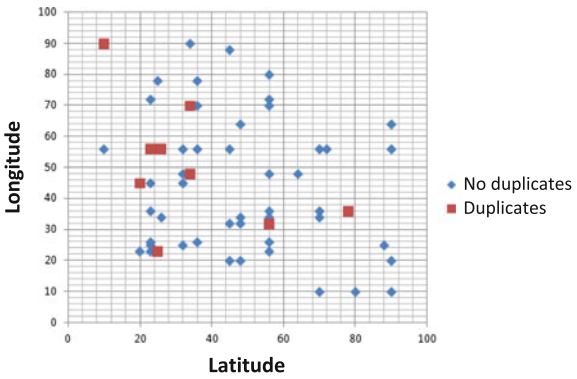


Fig. 2 Visualization of dataset

The dataset has been depicted graphically to visualize the distribution of different objects according to geospatial locations. The bullets marked as red show the coordinates where more than one object is found. This kind of data is not ignored and handled efficiently in k-dLst tree. All data is inserted in k-dLst tree structure using algorithm INSERT_k-dLst tree successfully (Fig. 2).

Now when we opt for searching of some data at particular location using algorithm SEARCH_k-dLst algorithm, it is capable to show all records related to that location, if exist.

For example, when a user gives the query like

Query: List all object_id lying at geospatial coordinates (10, 90), it will first search for location (10, 90) in k-dLst tree and then will traverse through all dataNode linked list in kdLstNode related to key (10, 90). The result of query will list all three objects with object_id 1, 19, and 66.

drksujatha23@gmail.com

Table 2 Query result—objects at location (10, 90)

Latitude	Longitude	Object_Id
10	90	Obj_Id 1
10	90	Obj_Id 19
10	90	Obj_Id 66

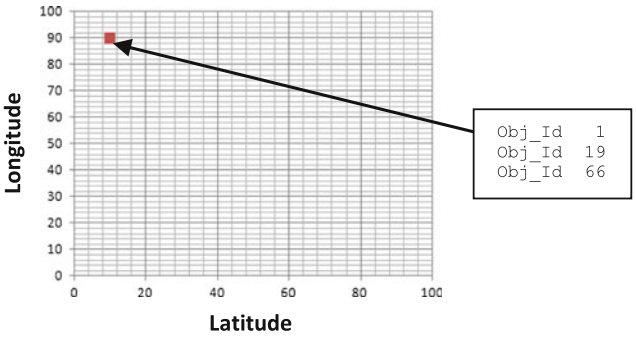


Fig. 3 Visualization of query result

Table 2 shows the records in dataset that are related to a particular geospatial location (10, 90). There are three records related to the same location. Figure 3 visualizes the result of Query 01.

7 Applications

There are lots of applications which need to store and retrieve spatial data. We require some efficient spatial indexing techniques to retrieve the required information fast. The applications where we can use some kind of spatial indexing include Location-based services, Fleet management, Sensor networks, Multimedia, Spatial Classification, Spatial clustering, Surveillance, Trajectory clustering, Rural/Urban land use, and many more.

8 Conclusion

This research paper has focused on a renowned data structure for indexing multi-dimensional points which is k-d tree. k-d trees are comparatively uncomplicated to understand and put into practice, and they are also useful for several important queries involving multidimensional keys, like point queries, range searches, and

drksujatha23@gmail.com

nearest neighbor searches. The research paper has listed the issue related to the handling of duplicate keys in spatial datasets. k-dLst tree is the combination of two structures, i.e., k-d tree and linked list. It can index duplicate key datasets efficiently.

References

1. Otair M (2013) Approximate K-nearest neighbour based spatial clustering using K-D tree. Int J Database Manag Syst (IJDMS) 5(1)
2. Bentley JL (1975) Multidimensional binary search trees used for associative searching. commun ACM 18:509–517
3. Crespo MM (2010) Design, analysis and implementation of new variants of Kd-trees. Master thesis, Departament de Llenguatges i Sistemes Informatics, Universitat Politecnica de Catalunya
4. Brown RA (2015) Building a balanced k-d tree in O(kn log n) time. J Comput Graph Tech (JCGT) 4(1):50–68
5. Friedman JH, Bentley JL, Finkel RA (1977) ACM Trans Math Softw 209–226
6. https://www.cs.cmu.edu/∼ckingsf/bioinfo-lectures/kdtrees.pdf (n.d.)
7. https://www.cs.umd.edu/class/spring2002/cmsc420-0401/pbasic.pdf (n.d.)

drksujatha23@gmail.com

Feature Extraction in Geospatio-temporal Satellite Data for Vegetation Monitoring



Hemlata Goyal, Nisheeth Joshi and Chilka Sharma

Abstract Today, geospatial technology becomes indispensable because of the technological advancement in automated data acquisition and the rapid growth of data, information and communication in 24×7 has been generating voluminous spatio-temporal data of earth surface. Voluminous spatio-temporal data is required to extract the features as several unrelated and redundant features, which may degrade the performance and fallout in extensive computation course of vegetation monitoring. The aim of this research is to select significant, efficient and effective selected features in monitoring of vegetation for Rajasthan state with standard feature extraction method of Principal Component Analysis (PCA), correlation and Spatial autocorrelation. In this paper, it applied these feature extractions on hydro-meteorological rainfall data, vegetation drought index as Standardized Precipitation Index (SPI) data and vegetation indices as Vegetation Condition Index (VCI) data to reduce the features related to vegetation monitoring.

Keywords Geospatial · PCA · Correlation · Spatial autocorrelation
VCI · SPI

1 Introduction

Geospatial technology responds to strengthening agriculture as it has the potential to extract out useful agriculture state, patterns, correlations in remotely sensed data, association of remotely sensed vegetation indices and rainfall, future prediction of

H. Goyal (✉)
Manipal University, Jaipur, India
e-mail: hemlata.goyal@jaipur.manipal.edu

N. Joshi · C. Sharma
Banasthali Vidyapith, Banasthali, India
e-mail: jnisheeth@banasthali.in

C. Sharma
e-mail: chilkasharma@gmail.com

© Springer Nature Singapore Pte Ltd. 2019

V. S. Rathore et al. (eds.), *Emerging Trends in Expert Applications and Security*,
Advances in Intelligent Systems and Computing 841,
https://doi.org/10.1007/978-981-13-2285-3_22

drksujatha23@gmail.com

agriculture and improving the knowledge of the agriculture domain [1]. Just of increasing powerful remote sensors, more computing power and enhancement in geospatial technologies themselves, it is a developing platform of selection for integrating and analysing massive amount of earth data [2]. Time series rainfall dataset for the months of June–September from 1981 to 2016 has been used to compute Standardized Precipitation Index (SPI) and multi-date NOAA-AVHRR NDVI based Vegetation Condition Index (VCI) for the months of July–October from 1997 to 2016. The results make out patterns of vegetation status coupled with feature extraction techniques such as PCA, correlation and spatial autocorrelation that gives effective, efficient and significant input features for vegetation monitoring. PCA has been used to identify spatio-temporal variability in multivariate time series data. Grouping has been performed, based on spatial autocorrelation in individual feature set and statistical correlation in between feature sets of Rainfall-SPI, Rainfall-VCI, Rainfall-SPI-VCI and Rainfall-VCI-SPI.

2 Related Work

Feature extraction is the necessary and efficient phase of successful multi-high dimensional spatial data mining application [3]. Minimizing the attributes, leading enhanced comprehensible model and simplify the practice of visualization techniques. PCA, correlation and spatial autocorrelation are well known and efficient feature extraction methods for time series dataset.

PCA is widely used in remote sensing to identify spatio-temporal variability in multivariate time series data [4]. It is a transformation method based on correlation, having constrained on the spatial and temporal pattern to be orthogonal [5]. It can only detain maximum variances in a fixed numbers of orthogonal components based on eigen analysis of the data correlation matrix [6].

In spatial data mining as the dimension of the data rise, the amount of data required to provide a consistent analysis increase exponentially [7]. Correlation-based feature selection follows the principal that "a good feature subset is one that contains features highly correlated with the class yet uncorrelated with each other [8]." It evaluates a subset by considering the predictive ability of each one of its features individually and also their degree of redundancy (or correlation).

Spatial autocorrelation in geospatial analysis helps to identify the degree of similarity of one object to other nearby objects and measure the correlation of a variable with itself (oneself) through space [9]. A variable has considered spatially

drksujatha23@gmail.com

correlated when there is an identical or a set pattern in its spatial distribution. 'Everything is related to everything else, but near things are more related than distant things (Geographer Waldo R. Tobler's, the first law of geography)'. **Moran's I is given by**

$$I = \frac{N \sum_i \sum_j W_{i,j}(X_i - \bar{X})(X_j - \bar{X})}{\left(\sum_i \sum_j W_{i,j}\right) \sum_i (X_i - \bar{X})^2} \quad (1)$$

In Eq. (1), N—no. of events; X_i—variable's value at a particular location; X_j — variable's value at different location; X—mean of the variable; W_{ij}—weight applied to the comparison between location *i* and *j*.

3 Study Region and Data

Rajasthan has been selected as the study area due to high variability of rainfall and vegetation. The NOAA-AVHRR smoothed fortnightly .tiff format VCI data, provided by NOAA NESDIS Center for Satellite Applications and Research (STAR) was downloaded at 4 km spatial resolution from the website https://www.star.nesdis.noaa.gov for July–October (1997–2016). Monthly average rain gauge station-wise rainfall data were obtained from the website (http://waterresources.rajasthan.gov.in) for the period of 1997–2016. Table 1 depicts the features used in this paper.

4 Research Methodology

Extraction of proper features is an important step in the spatial data mining process. There are 240 features of rainfall, SPI and VCI structures for 250 rain-gauge stations for Rajasthan state. In order to reduce the number of useless features and redundancy of time series dataset PCA, correlation and spatial autocorrelation feature extractions techniques has been used as given in Fig. 1.

Table 1 Description of the features

Feature	Description	Time range
VCI	Vegetation Condition Index derived from NDVI (Normalized Difference Vegetation Index)	July-October
Rainfall	Rain-gauge stationwise rainfall data	June-September
SPI	Standardized Precipitation Index (drought index) derived from rainfall	June-September

drksujatha23@gmail.com

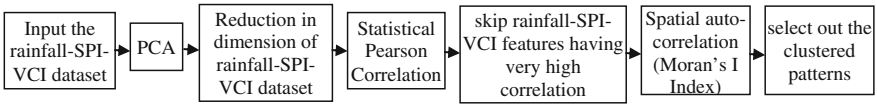


Fig. 1 Methodology of feature extraction in geospatio-temporal satellite data

4.1 PCA (Principal Component Analysis)

PCA is used to identify a feature set of Rainfall-SPI-VCI-derived variables that gives the importance of each component in terms of variance and correlation. The prcomp () function is used to generate the principal components in RStudio 3.3.2, displayed in the text view and the relative importance of the components is plotted [10]. Interpretability may reduce through using the derived variables that contribute most to the first few principal components rather than the original variables.

4.2 Correlation

Grouping is based on statistical correlation in between feature sets of Rainfall-SPI, Rainfall-VCI, Rainfall-SPI-VCI with statistical Pearson correlation, the entire dataset of Rainfall, SPI and VCI of 4 months for the period of 1997–2016 has been carried out to explore the correlation in feature extraction. A pair-wise correlation in between each feature has been computed and the graphic plot displays ellipse and colour to indicate the strength of any correlation.

4.3 Spatial Autocorrelation: Global Moran's I Function

'Global Moran's I' spatial autocorrelation function has been applied in SAGA GIS 2.1.2 environment in between SPIs at different rain-gauge stations to calculate that whether the pattern is random, clustered or dispersed, for a given set of features and related attributes. The obtained value, closer to +1.0 is considered as clustering and if it is closer to −1.0, then it is considered as dispersion.

5 Result and Discussions

In order to extract the features, PCA has been used to reduce the huge time series dataset for the period of 1997–2016. It selects the time series data with fewer dimensions and useful in determining patterns of association across variables. It

drksujatha23@gmail.com

performs an eigen value decomposition of the correlation matrix to find the principal axes of the shape formed by the scatter plot of the data. The eigen vectors represent the direction of one of these principal axes, and the eigen values are the variances of the associated component factors. The PCA results and statistics summary is shown in Table 2.

The screeplot of PCA in terms of variance, is shown in Fig. 2. It is evident with an elbow-shaped curve that first three components having maximum variance as compared to remaining components.

A time series database of rainfall, SPI and VCI dataset for each rain gauge weather monitoring station of Rajasthan has been created to calculate the correlation in between rainfall-SPI-VCI of 4 months (1997–2016). A pair-wise correlation in between each feature has been computed and shown as graphic plot in Fig. 3. The graphic plot displays ellipse and their colour are used to indicate the strength of any correlation. The result depicted in Fig. 3 concluded that mostly dataset retained blue ellipse instead of red, meaning thereby that there is positive correlation. Empty cells represent that there is no any correlation due to also back-gap months.

Table 3 shows the correlation between the datasets for Rajasthan. It has been established as the correlation between Rainfall-SPI, and processed SPI-VCI is very high. The computed correlation between Rainfall-SPI, Rainfall-VCI and SPI-VCI are 0.959, 0.644 and 0.673, respectively.

The correlation between SPI and VCI is depicted for all districts of Rajasthan state in Fig. 4. Correlation between SPI and VCI have been achieved to be negative in southern region including some parts of Sri Ganganagar and Bikaner due to presence of Indira Gandhi Canal, and because of Chambal river in Kota and surrounding districts. Some random parts of the state found negative correlation, e.g. Barmer, where negative correlation could have been due to forest and wasteland cover classes or may have been due to lag effect. The lag effect has been also taken into consideration as one month in between SPI and VCI to draw the correlation, but still the correlation was found to be negative. Thus, it has been concluded that a negative correlation is found between SPI and VCI in the area of forest, and wasteland cover, while in other parts of the state, positive correlation is achieved.

It has been an attempt to compute the spatial autocorrelations between the Rainfall, observed SP, and NDVI-based VCI values for 253 rain gauge stations correspondingly for June–September months.

Moran's I classified the Rainfall, SPI and VCI results as positive, negative and no spatial autocorrelation. It calculates that whether the pattern is random, clustered or dispersed, for a given set of features and related attributes. The obtained value, if closer to +1.0 is considered as clustering and if it is closer to −1.0, then it is considered as dispersion. Figure 5 is evident that all rainfall, SPI and VCI data have fallen out in cluster and random category, means data is auto correlated itself.

drksujatha23@gmail.com

Table 2 Statistics scored with PCA

PCA components	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Standard deviation	2.34	1.83	1.268	1.164	1.07	1	0.946	0.869	0.848	0.772	0.735	0.695	0.634	0.588	0.56	0.07	0.056	0.055	0.049	0.039
Proportion of variance	0.288	0.176	0.085	0.071	0.06	0.053	0.047	0.04	0.038	0.031	0.028	0.025	0.021	0.018	0.017	0	0	0	0	0
Cumulative proportion	0.288	0.465	0.549	0.62	0.681	0.734	0.781	0.82	0.858	0.89	0.918	0.943	0.965	0.983	0.999	1	1	1	1	1

drksujatha23@gmail.com

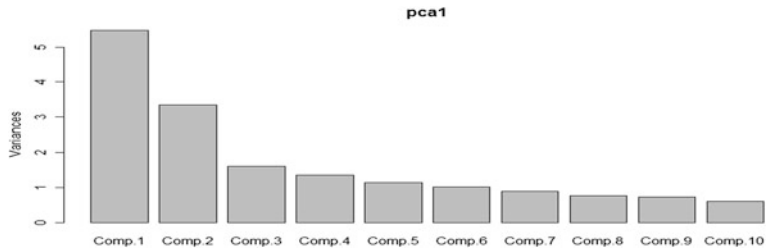


Fig. 2 PCA variance plot

It obtains the standard Z scored value, which shows that the clustering and dispersion are significant or has resulted due to a random chance. The observed rainfall corresponding to June, July, August and September months, the SPI corresponding to June–September and the VCI corresponding to July to October monsoon season month are critical to determine the agricultural phenomenon, as analysed for their spatial correlation's patterns. The obtained outcomes are depicted in Table 4.

Therefore, it is clearly observed in Table 4 that the separation of the rainfall values and distribution trend them of corresponding to each rain gauge stations of

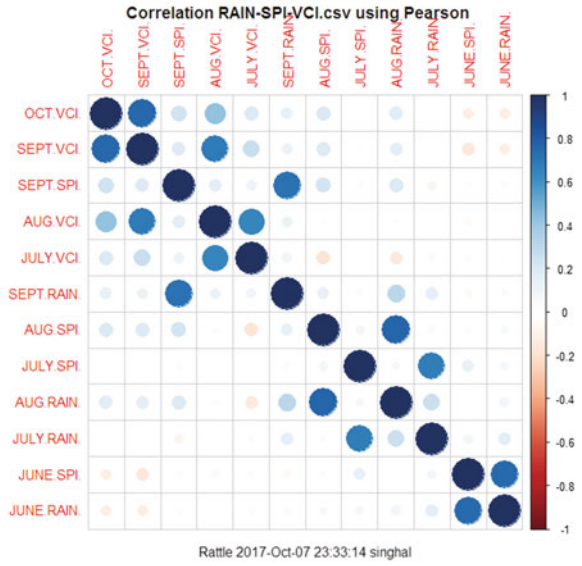


Fig. 3 Correlation in Rainfall-SPI-VCI for the period of 1997–2016

drksujatha23@gmail.com

Table 3 Correlation between the dataset

Process input	SPI				VCI				SPI and VCI			
Raw input	June	July	Aug	Sep	July	Aug	Sep	Oct	June–July	July–Aug	Aug–Sep	Sep–Oct
Rainfall	0.971	0.909	0.957	0.998	0.623	0.665	0.689	0.598	0.612	0.681	0.703	0.695
Average	0.959				0.644				0.673			

drksujatha23@gmail.com

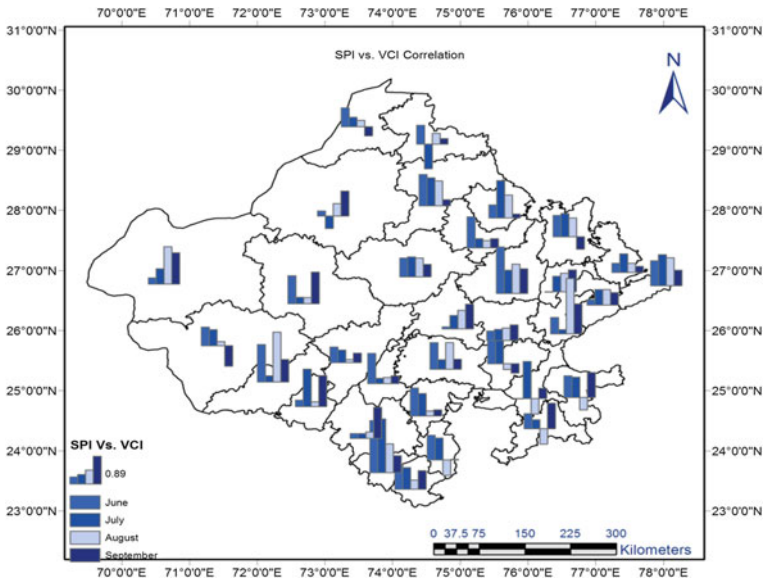


Fig. 4 District-wise correlation between SPI and VCI of June–September months

Rajasthan state have been determined that rainfall value corresponding to August month is statistically highly significant those of July month, and September than August month. It is concluded that VCI values of September month is statistically highly significant than July and August months. It is evident that the correlation between the observed rainfall values of June and July months are highly correlated as compared to August month, depicted by observing variance corresponding to each month as shown in Fig. 6. Similarly, the correlation between the observed SPI values of June and July months are very high as compared to August and the

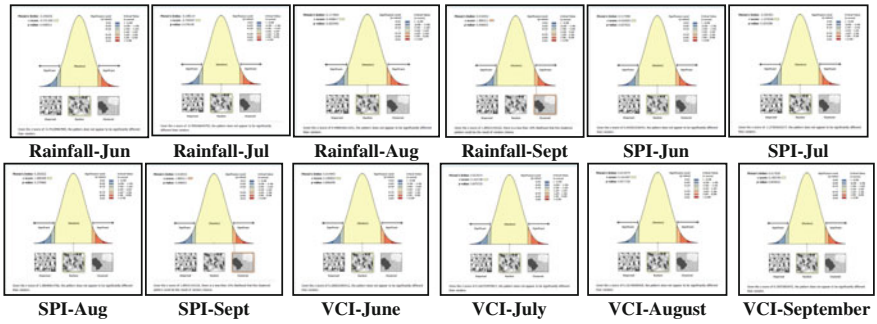


Fig. 5 Rainfall, SPI and VCI spatial autocorrelation (Moran's I) pattern

drksujatha23@gmail.com

Table 4 Spatial autocorrelation of rainfall, SPI and VCI by Moran's I

Spatial autocorrelation	Months	Moran's index	Expected index	Variance	Z-score	P-value
Rainfall	June	−0.206351	−0.004292	0.0743	−0.7413	0.45851
	July	0.14945	−0.004292	0.01132	0.18083	0.8565
	August	0.1176	−0.004292	0.07373	0.44892	0.65349
	September	0.513474	−0.004292	0.01211	0.16147	0.87173
SPI	June	0.117494	−0.004292	0.0756	0.44292	0.65782
	July	−0.354551	−0.004292	0.07546	−1.275	0.2023
	August	0.292622	−0.004292	0.07552	1.08041	0.27996
	September	0.516422	−0.004292	0.07548	1.89531	0.05805
VCI	July	0.014945	−0.004292	0.01132	0.18083	0.8565
	August	0.013414	−0.004292	0.01184	0.24292	0.87073
	September	0.013474	−0.004292	0.01211	0.26147	0.87173
	October	0.017635	0.004292	0.01255	0.17274	0.84482

correlation between the observed VCI values of July to October months is high as depicted observed variance value corresponding to each month for VCI.

6 Conclusion and Future Scope

It has been concluded, in order to extract the features with PCA that first three components having maximum variance as compared to remaining components, means significant and cannot be neglected. With statistical correlation, the correlation in between rainfall-SPI and processed SPI-VCI are very high. The computed correlation between Rainfall-SPI, Rainfall-VCI and SPI-VCI are 0.959, 0.644 and 0.673, respectively. With Pearson correlation, mostly dataset retained blue ellipse instead of red, meaning thereby that there is positive correlation. With spatial autocorrelation, in rainfall, SPI and VCI, all the dataset fall out in clustered and random category, which is the indication to itself, auto correlated. In future, this PCA can be extended with ICA (Independent Component Analysis).

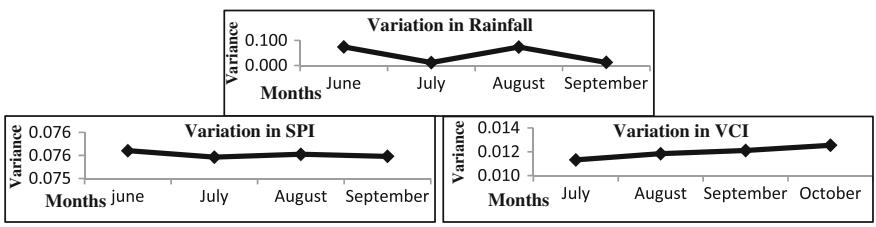


Fig. 6 Variation in rainfall, SPI and VCI

drksujatha23@gmail.com

References

1. Sharma L, Mehta N (2012) Data mining techniques: a tool for knowledge management system in agriculture. Int J Sci Technol Res 1(5):67–73
2. Manjula A, Narsimha G (2015) XCYPF: a flexible and extensible framework for agricultural crop yield prediction. In: IEEE 9th international conference on intelligent systems and control, pp 1–5
3. Liu H, Motoda H, Setiono R et al (2010) Feature selection: an ever evolving frontier in data mining, pp 4–13
4. De Almeida TIR, Penatti NC et al (2015) Principal component analysis applied to a time series of MODIS images: the spatio-temporal variability of the Pantanal wetland, Brazil. Wetl Ecol Manag 23(4):737–748
5. Wu JP, Wei S (1989) Time series analysis. Human Science and Technology Press, ChangSha
6. Verbesselt J, Hyndman R, Newnham G et al (2010) Detecting trend and seasonal changes in satellite image time series. Remote Sens Environ 114(1):106–115
7. Beniwal S, Arora J (2012) Classification and feature selection techniques in data mining. Int J Eng Res Technol (IJERT) 1(6)
8. Hall MA (2000) Correlation-based feature selection for discrete and numeric class machine learning. In: Proceedings of the 17th international conference on machine learning, pp 359–366
9. Bonham-Carter GF (2014) Geographic information systems for geoscientists: modelling with GIS, vol 13. Elsevier
10. Venables WN, Ripley BD (2002) Modern applied statistics with S. Springer

drksujatha23@gmail.com

Multiple Objects Tracking Under Occlusion Detection in Video Sequences



Sanjay Gaur, Sheshang Degadwala and Arpana Mahajan

Abstract This research displays an ongoing framework to identification of various occluding questions in element scenes. Article identification is a workstation engineering that bargains for identifying examples for same questions of a part (likely similar as peoples, vehicles, or buildings) for advanced pictures Also features. In the primary objective from claiming impediment identification from feature In utilizing Gaussian mixture model (GMM) strategy which will be foundation demonstrating will be should yield reference model What's more this reference model is utilized within foundation subtraction done each feature grouping may be compared against those reference model will focus time permits variety. Then impediment identification In light of Questions pixels qualities.

1 Introduction

Item identification may be a vital part from a canny feature reconnaissance framework. This object identification will be roused toward various applications, for example, surveillance, feature conferencing, man-machine interfaces, and also sports upgrade. Exact Also ongoing article identification will extraordinarily enhance that execution about object recognition, action dissection also high-keyed off chance Comprehension.

Video inpainting is acreage of computer vision, which is used for article tracking, motion detection, article detection, stereo vision, video processing, and

S. Gaur
Jaipur Engineering College & Research Centre, Jaipur, India
e-mail: sanjay.gaur@gmail.com

S. Degadwala (✉)
Madhav University, Sirohi, India
e-mail: sheshang13@gmail.com

A. Mahajan
Sigma Institute of Engineering, Vadodara, India
e-mail: mahajan.arpana@yahoo.com

© Springer Nature Singapore Pte Ltd. 2019

V. S. Rathore et al. (eds.), *Emerging Trends in Expert Applications and Security*,
Advances in Intelligent Systems and Computing 841,
https://doi.org/10.1007/978-981-13-2285-3_23

drksujatha23@gmail.com

video analysis. Occlusion botheration of accompanying article apprehension back one or added altar abutting to anniversary added in video orders. Here, consider attention to part of twins added altar choke anniversary, whether it is incompletely or wholly. Letter that these table to be adamant (Example. Vehicles) or disfigure (Example. People). It will be additionally anchored (Example. a pillar) and movable, in the part of situation it will be anchored or in gesture. Gabriel et al. [1] for appraise and analyze competences for the assorted video following schemes developed not for any to assorted requests, we accept begin it advantageous to advance academic ideas of matters, collections of bench and obstructions.

An overview of the system is shown in Fig. 1. At the start of input as video sequences, the video is converted into number of frames. When two or more objects is nearest to each other then find location of each object. The GMM methods are used for object detection. Stirring objects video sequence are subtraction using Gaussian approach in the ruining frames and its previous frame. Two or more objects are nearest to each other calculation based on threshold T value so generate the output of occlusion detection based on the background subtraction Gaussian mixture model (GMM) methods.

As shown in Fig. 2, item identification may be a workstation innovation organization that bargains with identifying instances about particular questions of a sure

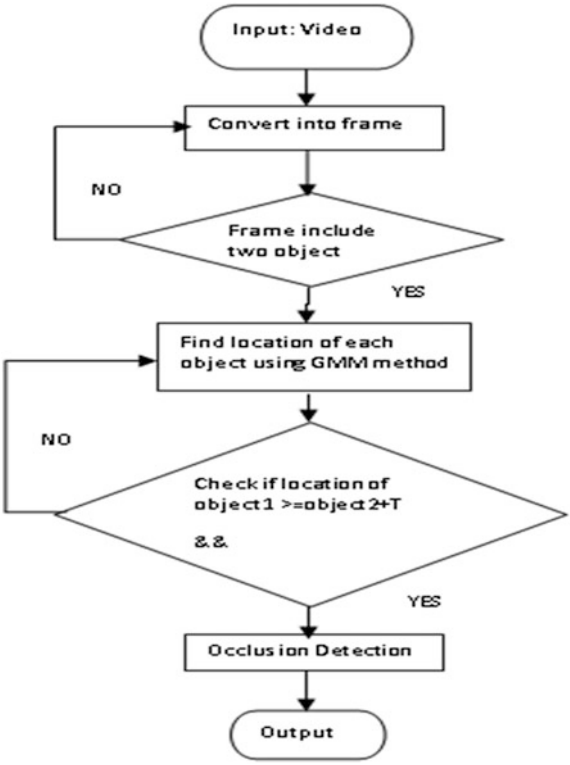


Fig. 1 Overview of approach [1]

drksujatha23@gmail.com

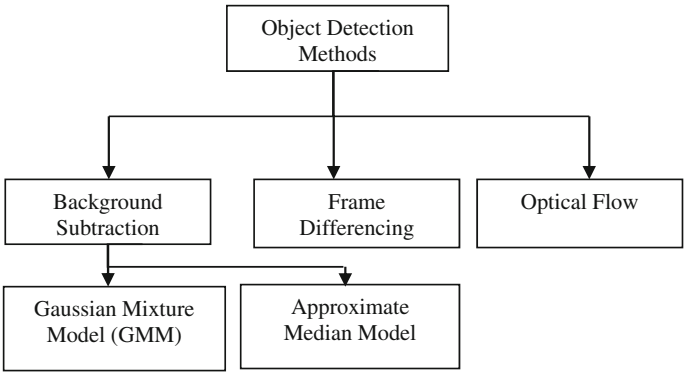


Fig. 2 Types of object detection methods

bunch (such as humans, buildings, or cars) clinched alongside advanced pictures Furthermore features. Object identification might be done by utilizing exactly essential strategies such as foundation subtraction [2], outline differencing [3] Furthermore optical stream [4].

2 Related Work

i. Frame differencing [3]: That vicinity about moving questions will be confirmed toward ascertaining the divergence between two exchange frames. Its count will be straightforward should actualize all the. It need An solid adaptability, for an assortment of element environments, Anyhow it is for the most part was troublesome with get finish framework for moving object, mindful will show up the spotless situation, Similarly as an aftereffect the identification about moving article will be not impeccable.
ii. Optical Flow [4]: Optical stream strategy will be should ascertain those outline optical stream field, Furthermore do grouping preparing as stated by the optical stream conveyance qualities from claiming picture. To move forward background, this strategy camwood get those supreme development data furthermore identify the moving object. Still, an extensive amount from claiming calculation, affectability to noise, lessened against commotion performance, make it not fitting for ongoing testing events.
iii. Background subtraction [2]: Foundation subtraction will be foundation displaying. It may be the center about foundation subtraction calculation. Foundation displaying must touchy sufficient to recognizing moving Questions. Foundation displaying may be should yield reference model. Clinched alongside foundation subtraction, each feature arrangement will be compared by the reference model on figure out conceivable variety. Those difference between current frames and the orientation plan By way of image

drksujatha23@gmail.com

part offer imperativeness on attendance from claiming moving Questions. The basis deduction system is to apply for difference system for the current outline and for basis span will recognize moving objects, focus recognitions yet altogether touchy of the transforms in the outside nature's domain.

3 Background Subtraction

i. Gaussian mixture Model (GMM) [5]: Show, Fig. 3, Gaussian mixture model (GMM) may be an parametric likelihood thickness work spoken to Likewise a weighted aggregate about Gaussian part densities. GMMs are parametric model of the likelihood conveyance of nonstop ability alternately features in a biometric framework. It incorporates color based following about an object in feature. It may be huge should recognizing moving Questions from an arrangement of features frames. The Gaussian mixture model to foundation subtraction system in those span pixels need aid deleted starting with the needed feature. GMM may be sensitive of the Different changes, for example, illumination, beginning and ceasing from claiming moving Questions. Perception is those observing of the behavior, movements' alternately different evolving majority of the data typically about individuals What's more every now and again previously, a surreptitious way [5].

ii. Approximate Median Filter [6]: Hint at Previously, Fig. 4, done estimated average channel technique recognizing past n frames of the feature are buffered, also average channel about cushion frames may be utilized for count about foundation. Those pixels for higher quality over the comparing foundation pixel esteem then it increase foundation pixel Toward 1. Similarly, to current pixel Hosting easier values afterward foundation pixel. It decrements foundation pixel toward one. With the goal likewise will bring half of the information pixels would more stupendous over the foundation and a large portion for short of what Eventually Tom's perusing the foundation utilizing. Estimated average technique [6].

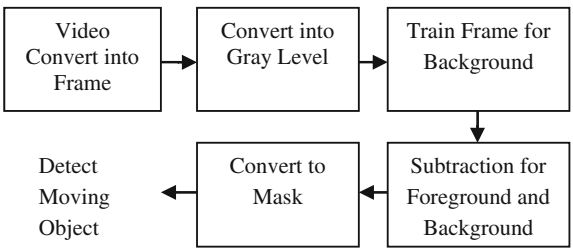


Fig. 3 Steps for Gaussian mixture model (GMM)

drksujatha23@gmail.com

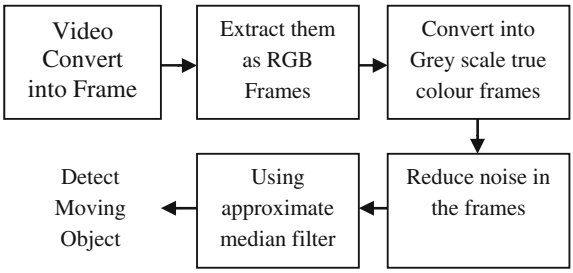


Fig. 4 Steps for approximate median filter

Eventually, Tom's perusing item identification outcome at both pixel also span levels, the foundation upgrade technique administers a suitableness foundation model under distinctive states. Demonstrate over Fig. 5.

For foundation subtraction step, each feature span will be compared for those reference foundation perfect; present pixel outline for veer off fundamentally to the foundation will a chance to distinguish. Then that an extent channel will be used to uproot little segments and affecting item locations will make picked up Furthermore changed of the unique subtraction picture with get the exact last division comes about. Recognizing the foundation mess and the comparability of the closer view district and the contextual, noises spots for extensive scope are unmoving occur afterward those morphological tenet sifting. With tackle to difficult, we investigate the right communication times about each spot inconsecutive edges, accordingly dubious clatters will make uprooted.

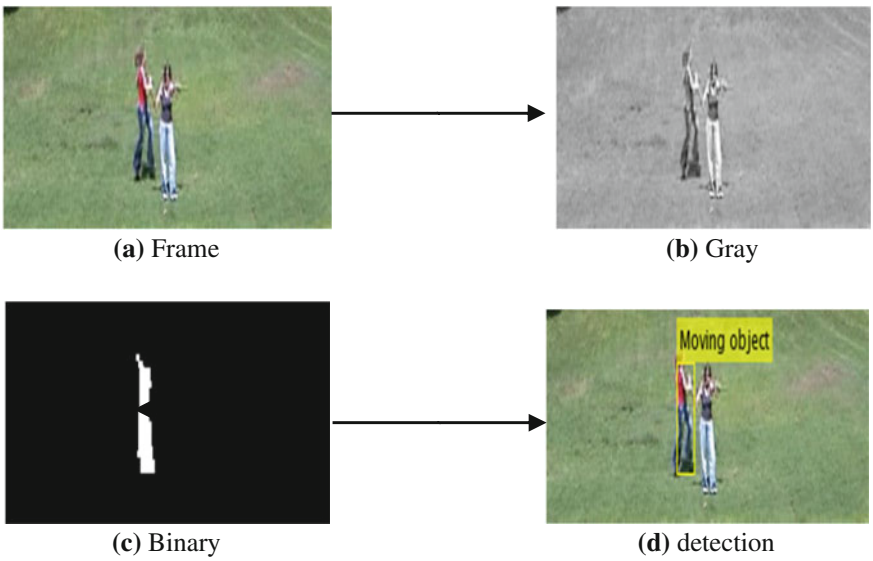


Fig. 5 Moving object detection. **a** Video as input. **b** Gray level. **c** Filtering with morphology result. **d** Result object detection

drksujatha23@gmail.com

4 Occlusion Detection

When there is more than one person, we just have to keep detection of which subset of frames belongs to each moving object [7]. When complete or partial occlusion occurs, we do not delete the frames that are no longer visible. The occluded frame retains their statistics, while statistics of other frames are updated slowly. Upon reemerging of an occluded region, the classification is automatically correct, provided the location of the region has not changed significantly during the time when it is occluded. This is because the remerging pixel will still have the highest likelihood to being assigned to their correct frame rather than any of the other frames.

As shown in Fig. 6 Xiao et al. [8] tells the two sorts of impediment going on previously, frames. Those main body of evidence is movement occlusion, the place the impediment era will be because of item movement and the blocked regions starting with two frames need aid not overlapped toward those same area. The second situation will be mismatching the place the blocked locales from diverse pictures need aid overlapped in those same location. The not matching part might occur under diverse circumstances, for example, object seeming/vanishing, gumshoe, shade variation, alternately extensive item deformity (shrinking or expanding), and so on. Should identify such occlusion, restricted is checking those consistency the middle of those ahead Furthermore retrograde stream. Assuming that the regressive and forward stream will be constant, the pixel will be recognized concerning illustration not on two object occlusion. Though, this forward–backward corresponding might not brand reliable to certain bags, for example, not same part the place the stream inside the both covering blocked areas might be zero.

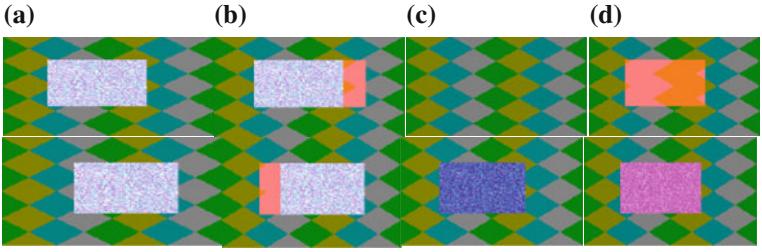


Fig. 6 Those the event of movement blocking, the place a box is touching starting with those port (the uppermost point frame) of the correct lateral (lowest frame). **b** Relating blocked regions about **a** are cover of pregnancy On red and the blocked regions spot toward separate positions because of those object's movement. **c** The body of evidence of mismatching, the place the highest point may be those to begin with span What's more a rectangle abruptly gives the idea in the second span (the base one). **d** Those comparing blocked territories from claiming **c** would likewise cover of pregnancy On red, Anyway in this the event, these blocked locales would overlapped during the same area

drksujatha23@gmail.com

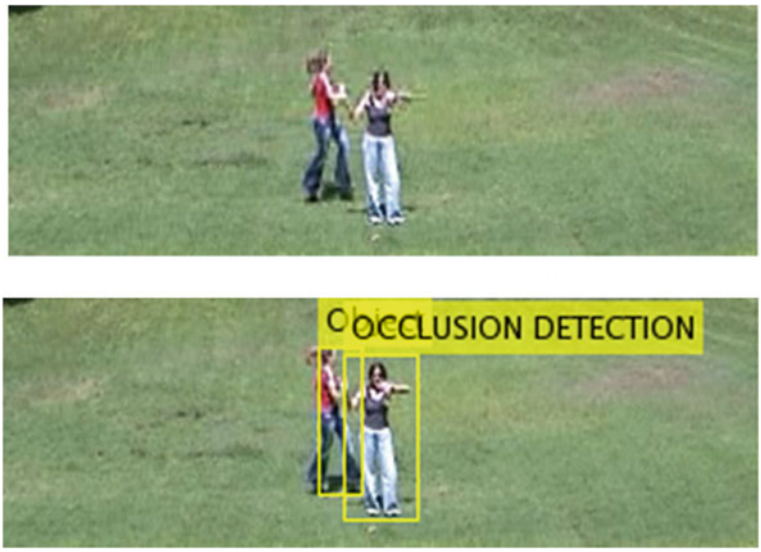


Fig. 7 An order of two interrelating person's detection with weighty blocking in an outdoor situation

Indicate clinched alongside Fig. 7 indicates a sample for identification two cooperating persons in open air nature's domain. Those yellow box indicates the position of the pernickety for impediment.

5 Analysis

See Table 1.

Table 1 Analysis of detection method

Method	TP	TN	FP	FN	TG	TF	Accuracy (%)
Approximate median	90	55	0	55	200	200	72.5
Adaptive Gaussian mixer model (GMM)	115	50	0	35	200	200	82.5

drksujatha23@gmail.com

6 Conclusion

From various studies of different papers related to "occlusion detection". We have presented the proposed approach for dynamic background object detection while occlusion. For that object detection methods for inpainting is used such as frame differencing method, optical flow method, Gaussian mixture model (GMM) method, approximate median method. It can be concluded that GMM method is best for good quality result, noise remove, and handle illumination when compared to approximate median method. We observe that in video sequences and location of object, occurrence of occlusion. This approach identifies object during occlusion.

References

1. Gabriel PF, Verly JG, Piater JH, Genon A (2003) The state of the art in multiple object tracking under occlusion in video sequences. In: Proceedings of advanced concepts for intelligent vision systems, pp 166–173
2. Parekh HS, Thakore DG, Jaliya UK (2014) A survey on object detection and tracking methods. Int J Innov Res Comput Commun Eng (IJIRCCE) 2
3. Rakibe RS, Patil BD (2013) Background subtraction algorithm based human motion detection. Int J Sci Res Publ (2013)
4. Chauhan AK, Krishan P (2013) Moving object tracking using gaussian mixture model and optical flow. Int J Adv Res Comput Sci Softw Eng
5. Santosh DHH, Venkatesh P, Rao LN, Kumar NA (2013) Tracking multiple moving objects using Gauusian mixture model. Int J Soft Comput Eng (IJSCE) 3(2)
6. Rao GM, Satyanarayana C (2014) Object tracking system using approximate median filter, Kalman filter and dynamic template matching
7. Khan S, Shah M (2000) Tracking people in presence of occlusion. In: Asian conference on computer vision
8. Xiao J, Cheng H, Sawhney H, Rao C, Isnardi M (2006) Bilateral filtering-based optical flow estimation with occlusion detection. In: Proceedings of computer vision—ECCV 2006. Springer, Berlin, Heidelberg, pp 211–224
9. Piater JH, Crowley JL (2001) Multi-modal tracking of interacting targets using Gaussian approximations. In: Second IEEE international workshop on performance evaluation of tracking and surveillance, vol 14, p 58
10. Rosales R, Sclaroff S (1998) Improved tracking of multiple humans with trajectory prediction and occlusion modeling. In: IEEE CVPR workshop on the interpretation of visual motion
11. Yang T, Pan Q, Li J, Li SZ (2005) Real-time multiple objects tracking with occlusion handling in dynamic scenes. In: IEEE computer society conference on computer vision and pattern recognition, CVPR 2005, vol 1, pp 970–975. IEEE
12. Chang, T-H, Gong S, Ong E-J (2000) Tracking multiple people under occlusion using multiple cameras. In: Proceedings of BMVC, pp 1–10

drksujatha23@gmail.com

IoT Platform for Smart City: A Global Survey



Rakesh Roshan, Anukrati Sharma and O. P. Rishi

Abstract In today's scenario, Internet of Things (IoT) platform is the integrated part for the development of smart cities across the globe. The prediction of expert is that there will be a 30 billion object for Internet of things by 2020. So, there is a need of lots of research required for Internet of Things (IoT) in next 20 years. And we cannot imagine for the smart city without Internet of Things (IoT). In this paper, there will be the survey of the research which is already done in last 5 years than what to do in future for developing the better architecture of smart city. Multilevel architecture and framework of the smart city is also proposed in this paper. The suggested components of the smart cities are explained in the last section of the paper.

Keywords IoT · Internet of things · Smart city · Sensors

1 Introduction

"A smart city is a vision for urban development by integrating various Information and Communication technology (ICT) and solutions of IoT in a safe mold to deal with a city's components—the city's components includes, nearby department's data centers, transportation frameworks, schools, libraries, medical facility, power houses, water supply systems, law enforcement, Management of waste, and other services" [1].

Approximately, 31% of India's present population live in urban areas and contribute 63% of GDP. By 2030, the 40% of India's population live in urban

R. Roshan (✉) · A. Sharma · O. P. Rishi
University of Kota, Kota, India
e-mail: RRoshan1980@gmail.com

A. Sharma
e-mail: dr.anukratisharma@gmail.com

O. P. Rishi
e-mail: dr.oprishi@uok.ac.in

© Springer Nature Singapore Pte Ltd. 2019

V. S. Rathore et al. (eds.), *Emerging Trends in Expert Applications and Security*,
Advances in Intelligent Systems and Computing 841,
https://doi.org/10.1007/978-981-13-2285-3_24

drksujatha23@gmail.com

areas and contribute 75% of GDP. To accomplish, the objective, extensive improvements of physical, institutional, social, and financial foundation are required. All are essential in enhancing the personal satisfaction and pulling in individuals and venture, getting underway a righteous cycle of development and advancement. Improvement of smart cities is a stage toward that path.

The objective of smart cities may be

- To enhance effectiveness of open utility in water, gas, power supply transportation, communication, and in this manner understand an advanced way of life.
- To enhance livability index across the corridors of development which are required to push economic development of the cities.
- To provide safe and secure living environment using technological innovations for healthy growth of the smart cities.
- To intelligently use information technology to give the facility to migrant population with e-administration frameworks being the spine of foundation.
- To provide the platform for decision-makers and good practice owners together in e-government academies and international conferences.
- To provide robust IT connectivity and digitalization.
- Efficient cities mobility and open transport.

Top 5 smart cities of the world are (Table 1).

Table 1 Top 5 smart cities

Name of city	Why smart?
Seoul	Best heathcare facilities for the disabled and the elderly, for that they will be provided with second-hand smartphones and tablets to ensure medical attention when needed Ready for hosting 5G mobile technology
San Francisco	Technology used effectively for improvement in transport, energy, water supply, and waste management LED street lights, EV charging infrastructure, and parking sensors
Hong Kong	Mission "Digital 21 strategy" Excellent e-government services for its citizens Highest smartphone penetration in the world Contactless card payment in public transport
Singapore	Use of cameras GPS and sensors to prevent traffic congestion and also predict jams Innovative water management system
Rio de Janeiro	The center of operations set up by IBM in the city has linked all major departments of the city Discovery of offshore oil field

drksujatha23@gmail.com

2 Internet of Things and Their Characteristics

There are lots of research because IoT is still very young, and the proper definition of IoT is not available. Internet of Things is divided into three parts:

- Internet,
- Things (Sensors, home appliances, car, building, etc.),
- Software (The programs to manage the sensors, knowledge, etc.).

Also, IoT can be used for industrial application, smart homes, and smartphones. According to the community of RFID [2], the IoT can be defined as "Worldwide network of interconnected things uniquely addressable with the standard communication protocols." The Internet of Things gives the opportunity to connect the things to people anytime and anywhere. The different terms used by the different companies for IoT are as follows: IBM used "Smarter Planet", Cisco used "Internet of Everything", and GE used "Industrial Internet"; the main objective if these companies are to improve production time, reduce energy consumption, avoid accidents, and reduce downtime using different sensors. Components of IoT Infrastructures are

- Heterogenous devices,
- Constrained of the sensors/devices,
- Real-time interaction,
- Heterogeneous network,
- Huge number of network and devices, and
- Intelligence.

3 Review of Literature

As indicated by Zanella et al. [3], the IoT might have the capacity to coordinate straightforwardly and flawlessly an expansive number of various heterogeneous devices, while giving open access to chose subsets of information for the advancement of a vast size of computerized administrations. To develop a universal architecture of IoT for smart cities are extremely difficult task because of the huge variety of devices, protocols, and services those are associated with such a framework. Zanella et al. [3] give an extensive overview of the empowering technologies, architecture, and conventions (Protocols) for the urban IoT. They also presented and discussed the specialized solutions and best-rehearse guidelines utilized in the Padova Smart City project. Zanella et al. [3] discussed for service architecture that can extract from the particular attributes of the single advancements/technologies and provide combined access to the following services: architectural health of buildings, noise tracking, air pollution control, energy

drksujatha23@gmail.com

consumption by city, waste management system, traffic congestion, smart parking and lighting, automation and security of home/apartment/offices, etc.

Suciu and Vulpe [4] explored the characteristic of a cloud platform for smart cities deployment with the intention of validating the platform's capability to provide tailored IoT components via the cloud computing.

Navarro et al. [5] designed the new framework for energy efficient commercial transport for smart cities and applied in Spain. The vehicle division is in charge of 30% of the CO_2 emanations in EU, coming to up to 40% in urban regions. Navarro demonstrates the consequences of the live trial of smart city urban logistics solutions in the urban communities of Barcelona and Valencia that considered of joining the utilization if tricycles and transshipment terminals for the last-mile conveyance of packages and little shipments. A mindful investigation of the quantitative results of the pilot test in the two urban areas is displayed from alternate points of view: operational, financial, environment, social, and energy efficiency.

Djahel et al. [6] presented a thorough review covering various technologies utilized in various stages of traffic management system. They also discussed the potential utilization of smart vehicles and social networking sites to enable quick and exact detection of incidents and mitigation, and briefly presented current initiatives in Europe and worldwide to foster progress in smart transportation.

Misbahuddin et al. [7] approach traffic management issues from an IoT angle. In the practical situation of the holy city of Makkah, in Saudi Arabia, where traffic flow is very effective by the continuous visiting of pilgrims throughout the year, they suggested new and more robust controlling algorithms and strategies should be implemented. According to proposal given by authors, flow can be dynamically monitored and controlled in a number of ways, such as by onsite traffic officers through their smartphones, or centrally through the Internet.

Zheng et al. [8] presented a prediction system for the parking inhabitance rate using three attribute sets with chosen parameters to represent the utilities of these attributes. Their approach considered two situations in light of real-time car parking information system that has been gathered and dispersed by the City of Melbourne (Australia) and City of San Francisco (USA). Authors also analyzed the relative qualities of various machine learning techniques for prediction purposes, so as to improve the efficiency of parking management systems.

4 Observation of the Global Review

After the study of the implementation of some of the smart city, it is observed that the all technical components can be categorized into three parts:

- Intelligent infrastructure facility: Intelligent infrastructure means the physical infrastructure which contains integration or interconnected of sensors, networks, and computing hardware and software. This is the rapid evolution in the traditional infrastructure of all types, such as home to bridges, streets, dustbins, and

drksujatha23@gmail.com

every small thing which is related to us. Here, name is intelligent infrastructure because obviously all will be connected to Internet to collect multidimensional data. These things make intelligent infrastructure more cost-effective. The most common thing in the intelligent infrastructure is smart objects and smart communication. These two things are the motivation for the development of intelligent infrastructure.

- Innovation of processes (Big Data): The huge amount of sensors will produce vast amounts of information from the infrastructure of smart cities and send/receive messages to one another department. So, effective data management tools are required to handle or process all those huge amount of data. In today's market, "Big data" is used for huge amount of data and many tools are available to manage these data according to the requirement of the systems.
- Big Data is an idea to producing knowledge in which various innovative strategies are connected to the aggregated information, administration, and investigation of enormous and multidimensional information—information so huge, so changed, and examined at such speed that it surpasses the abilities of customary information administration and analysis tools.
- Smart energy management: Smart energy management is also required because of lots of energy consumption by the huge amount of sensors of the infrastructure. The smart energy management helps to improve the performance as well as low-energy consumption of the electricity. Smart grid is one of the concepts for the smart energy management. Smart grid infrastructure integrates the Information & Communication Technology (ICT) into the power supply network to detect and smartly work on continuous flow of information effectively and efficiently.

The designs of framework for smart cities are very difficult and complex. Also, there are lots of challenges due to rapid urbanization of India. The tentative designs of framework for smart cities of India are visualized in Fig. 1.

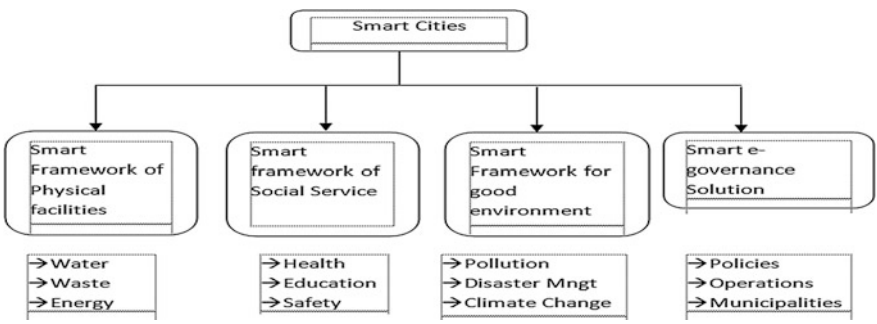


Fig. 1 Suggested components of smart cities

drksujatha23@gmail.com

5 Conclusion

In this, global survey for the Internet of Things platform of different smart cities tried to know the technical implementation of smart cities in different parts of the world. Development of smart cities is not a simple task; it requires the involvement of the people of the city or country. Also, this paper focussed on the various issues and challenges to implement the IoT in smart city according to geographical area, education, etc. The smart city is the concept which is based on an integration of various components. This is also the conclusion that Internet of Things is just a support system for the smart cities and cannot be replaced by the subsystems of the smart city. The IoT infrastructure can support and give the service to only one component but it can be used in every component of the smart cities implementation. The future challenges for the smart city are security, environment, and educate the people to adopt the smart city concept.

References

1. www.wikipedia.com
2. Atzori L, Iera A, Morabito G (2010) The internet of things: a survey. Comput Netw 54 (15):2787–2805
3. Zanella A, Bui N, Castellani A, Vangelista L, Zorzi M (2014) Internet of things for smart cities. IEEE Internet Things J 1(1):22–32
4. Suciu G, Vulpe A (2013) Cloud computing and internet of things for smart city deployments. In: International conference on challenges of the knowledge society (CKS) 2013, Romania, 17–18 May 2013, pp 1409–1416
5. Navarro C, Roca-Riu M, Furio S, Estrada M (2015) Designing new models for energy efficiency in urban freight transport for smart cities and its application to the Spanish case. In: 9th international conference on city logistics, Tenerife, Canary islands, Spain, 17–19 June 2015, pp 314–324
6. Djahel S, Doolan R, Muntean GM, Murphy J (2014) A communication-oriented perspective on traffic management systems for smart cities: challenges and innovative approaches. IEEE Commun Surv Tutor 17(1):125–151
7. Misbahuddin S, Zubairi JA, Saggaf A, Basuni J, Wadany SA, Al-Sofi A (2015) IoT based dynamic road traffic management for smart cities. In: 12th international conference on high capacity optical networks and enabling/emerging technologies (HONET), Islamabad, 21–23 Dec 2015, pp 1–5
8. Zheng Y, Rajasegarar S, Leckie C (2015) Parking availability prediction for sensor-enabled car parks in smart cities. In: 2015 IEEE tenth international conference on intelligent sensors, sensor networks and information processing (ISSNIP), Singapore, 7–9 April, 2015, pp 1–6

drksujatha23@gmail.com

Incessant Ridge Estimation Using RBCA Model



Sandeep Kumar Sharma, C. S. Lamba and Vijay Singh Rathore

Abstract This manuscript represents a process for improving the quality of an image ridge using cellular automata outline. Aim of the article is to estimate continuous and accurate ridges or improve the quality of estimated ridge that makes continuous and accurate ridge. The RBCA method used nearby neighbors with radius 1 or 2 to estimate extremely incessant ridges. RBCA produces threshold value of a pixel using all pixels up to radius 2. The proposed method used that value to estimate the targeted incessant ridge. During this process, image attributes as intensity, contrast, and brightness are also customized to accomplish the target. The investigational outcomes state competence of the proposed technique.

Keywords RBCA · Ridge · Cell

1 Introduction

Image appearances like edge, ridge, Blob, and intersect point share the similar properties but minor difference. An edge is a border where the intensity has been changed, and ridge is also an edge with thin and darker line or it can be formed with the points where the intensity gray level reaches a local acute in a given path. This article anticipated a novel method to estimate an edge that has better quality. The incessant ridge is a continuous, thin, and darker edge. The objective of Incessant Ridge Estimation (IRE) is to get improved the illustration fascia of processed image

S. K. Sharma (✉)
RTU, Kota, India
e-mail: mcasandy2006@gmail.com

C. S. Lamba
RTU DRC, Kota, India
e-mail: lamba5@rediffmail.com

V. S. Rathore
Jaipur Engineering College & Research Centre, Jaipur, Rajasthan, India
e-mail: vijaydiamond@gmail.com

© Springer Nature Singapore Pte Ltd. 2019

V. S. Rathore et al. (eds.), *Emerging Trends in Expert Applications and Security*,
Advances in Intelligent Systems and Computing 841,
https://doi.org/10.1007/978-981-13-2285-3_25

drksujatha23@gmail.com

or to exchange an input image into an appearance that any human or machine can easily analyze. The primary goal of IRE is to determine and modify current attributes of an image to compose it further pertinent for any obligation and a defined viewer.

The quality of an image edge is justified on the behalf of its pixels intensity, continuity of pixels, contrast, and brightness of pixels. The quality of a pixel can be improved using the neighbor pixels. All the neighbor pixels play an important role by improving quality of image edge [1]. For better quality, we can use all the possible nearest neighbor pixels. The proposed method focuses on all the neighbors up to radius 2 [2].

Cellular Automata (CA) is a group of grids in the form of cell, which has finite number of states. CA is also called rule-based automata machine which consists of M * N consecutive cell [3]. Each cell determines next state on behalf of the present state of neighbors using transition function. Transition function is a method which processes source image pixels and generates new image pixels. CA has many rules on behalf of total pixels. Radius-Based Cellular Automata (RBCA) is a type of CA's rule, which targets all the neighbors up to two outer layers. The entire CA cell represents the pixels of an input image. There are different methods formed for image edge detection like histogram methods [4], derivation-based methods [5], operators [6], and different types of wave-based methods [7]. However, the experimental result shows the efficiency of existing methods and proposed method.

The remaining manuscript is structured as follows. Section 2 represents structure and construction of CA that is used in anticipated method; Sect. 3 represents principles on proposed method works; Sect. 4 represents algorithms for solving problem; and Sect. 5 describes investigational outcome. Section 6 represents the final remarks.

2 CA's Structure and Construction

2.1 Structure

Cellular automata (CA) have a structure in the form of group of neighbors [8]. CA has five types of neighbors (Left, right, upper, below, and corner neighbors). These five types of neighbors called inner neighbor and the neighbors outside of all these neighbors called outer neighbors. The corner neighbors also classified into four corners such as upper left corner, lower left corner, upper right corner, and lower right corner. The four-neighbor structure (Left, right, upper, and below) is known as Von Neumann structure, and the eight-neighbor structure is known as Moore neighborhood structure [9]. The outer neighbor of all eight neighbors is called as outer totalistic cellular automata (CA) structure [10]. The CA's neighbor structure is illustrated in Fig. 1.

drksujatha23@gmail.com

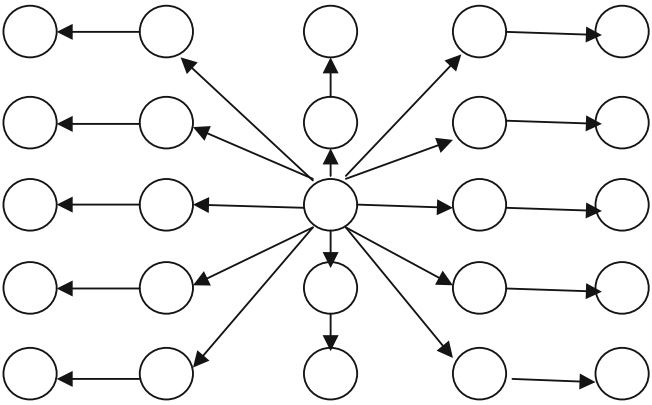


Fig. 1 CA's neighbor structure

2.2 Construction

Cellular automata (CA) [11] has extensive outline for various types of appliances. As discussed CA is a rule-based automata machine that can design number of rules based on neighbors [12]. For example, if a cell has n total number of neighbors than it can design 2^n number of rules. CA has a structure to represent all states of a cell called 5-tuple structure. The 5-tuple is represented in the form of < C, S, N, F, NS > where C is the current state, S is the group of predetermined states, N is the group of all nearest neighbors, and F represents a conversion function which consigns a novel status (N) to a pixel. CA neighbors classified as inner and outer neighbors based on current cell radius [13]. Von Neumann method embraces only four neighbors left, right, above, and below except corner neighbors; Moore neighborhood method embraces all the eight nearest neighbors up to radius 1, whereas radius-based cellular automata embrace all the nearest 24 neighbors up to radius 2. The entirety 8 and 24 of radius 1 and radius 2 gave by equation numbers 1 and 2 [14].

$$\mathrm{TOTAL_{NBR}} = \mathrm{LEFT_{NBR}} + \mathrm{RIGHT_{NBR}} + \mathrm{UPPER_{NBR}} + \mathrm{BELOW_{NBR}} \quad (1)$$

$$\mathrm{TOTAL_{NBR}} = \mathrm{LN_{NBR}} + \mathrm{RN_{NBR}} + \mathrm{UN_{NBR}} + \mathrm{BN_{BR}} + \mathrm{LC_{NBR}} + \mathrm{RC_{NBR}} + \mathrm{U_{NBR}} + \mathrm{B_{BR}} \quad (2)$$

where

LN = Left Neighbor
RN = Right Neighbor
UN = Upper Neighbor
BN = Below Neighbor
LC = Left Corner Neighbor
RC = Right Corner Neighbor
UC = Upper Corner Neighbor
BC = Below Corner Neighbor

drksujatha23@gmail.com

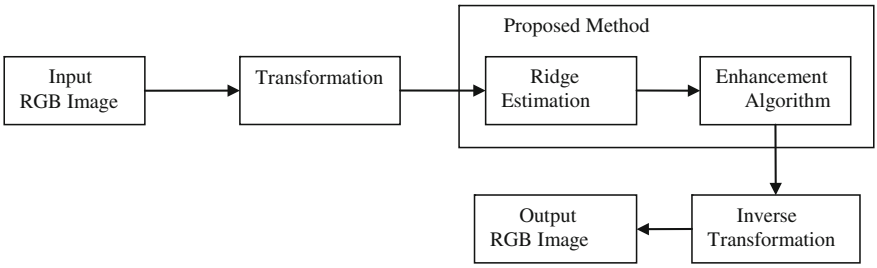


Fig. 2 Working principle

The entirety neighbors up to radius 2 given by Eq. 3.

$$TN_{i,j} = \sum_{k=i-2}^{i+2} \sum_{l=j-2}^{j+2} CC_{K,1} - \sum_{k=i-1}^{i+1} \sum_{l=j-1}^{j+1} CC_{k,1} \tag{3}$$

where TN is the total neighbors, CC is the current cell, and i, j, k, l indicate the positions of current cell.

3 RBCA Working Principle

Figure 2 shows the whole steps of research work. The total work divided into four steps. The first step is to acquisition of an image, the second stage translates the contribution image into gray image [15], the third step is the proposed work which estimates enhanced ridges, and the final stage renovates the outcome image into RGB form [16]; this is the final output of proposed research work.

4 Proposed Algorithm

To overcome with the objective should follow all the steps of working principal. Use MATLAB code for transformation and inverse transformation process. CA's structure and framework are used to estimate ridge and increase the quality of estimated ridge [17]. The anticipated RBCA technique is capable of exertion on all types of images. To accomplish this type of capability, transformation step is compulsory because CA can work only on cells or grids [18, 19]. The anticipated system has five steps for estimating final ridge of an image. The building block figure of image ridges assessment shows in Fig. 3.

drksujatha23@gmail.com

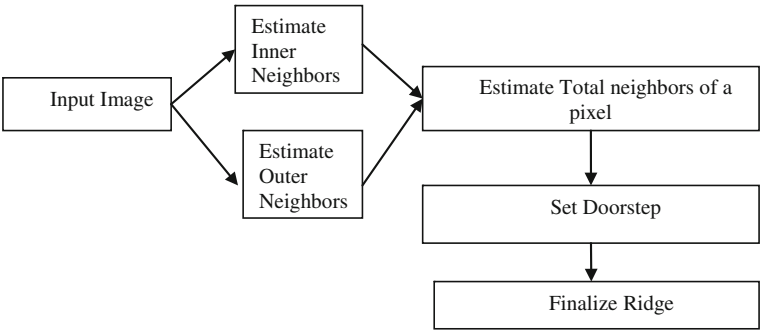


Fig. 3 Block diagram of RBCA method for ridge estimation

Let I(M, N) is the input image, INBR(i, j) is the inner neighbor, and ONBR(i, j) is the outer neighbor of a pixel PI(i, j).

Set total number of inner neighbors of pixel p(i, j).

$$\begin{aligned}\mathrm{INBR}(\mathrm{i},\ \mathrm{j}) = {} & \mathrm{PI}(\mathrm{i},\ \mathrm{j}-1)+\mathrm{PI}(\mathrm{i},\ \mathrm{j}+1)+\mathrm{PIp}(\mathrm{i}-1,\ \mathrm{j}-1)+\mathrm{PI}(\mathrm{i}-1,\ \mathrm{j}) \\ & +\mathrm{PI}(\mathrm{i}-1,\ \mathrm{j}+1)+\mathrm{PI}(\mathrm{i}+1,\ \mathrm{j}-1)+\mathrm{PI}(\mathrm{i}+1,\ \mathrm{j})+\mathrm{PI}(\mathrm{i}+1,\ \mathrm{j}+1)\end{aligned}$$

Set total number of inner neighbors of pixel p(i, j).

INBR (i, j) = PI(i, j-1) + PI(i, j+1) +PIp(i-1, j-1) + PI(i-1, j) + PI(i-1, j+1) + PI(i+1, j-1) + PI(i+1, j) + PI(i+1, j+1)

Set total number of outer neighbors of pixel p(i, j)

Step 1:- Repeat step 2 to 4 until i≤M
Step 2; - Repeat step 3 to 4 until j≤N
Step 3:- if(i = 1 || **i** = M)
if (j = 1 OR **j** = N)

Then
P=1; Q=1; R=1;
if (j = 2 OR **j** = N-1)
Then
P=1; Q=2; R=1;
else
Then
P=2; Q=2; R=1;
else if(i = 2 || **i** = M-1)
if (j = 1 OR **j** = N)
Then
P=1; Q=1; R=2;
if (j = 2 OR **j** = N-1)
Then
P=1; Q=2; R=2;
else
Then
P=3; Q=1; R=1;
else if(i = 3 OR **i** = M-2)
if (j = 1 OR **j** = N)
Then
P=2; Q=1; R=2;
if (j = 2 OR **j** = N-1)
Then
P=3; Q=1; R=1;
else
Then
P=4; Q=2; R=2;
else
Then
P=4; Q=2; R=2;
Step 4:- ONBR (i, j) = 3* P + Q + R;

drksujatha23@gmail.com

Estimate total neighbors

Step 1: Repeat step 2 to 3 until i $\leq$ M
Step 2: Repeat step 3 until j $\leq$ N
Step 3: TNBR (i, j) = INBR (i, j) + ONBR (i, j)

Set Threshold value

Let s1 and s2 are background and object classes of an input image. The total pixels in these classes are n1 and n1. The total pixels in input image are TN. The mean (μ) and variance (σ^2) of the classes are given by

$$\mu_i = \frac{\sum s_i}{n_i}$$

$$\sigma_i^2 = \frac{\sum (s_i - \mu_i)^2}{n_i}$$

$$T_i = \sum p_i \sigma_i^2$$

Where

$$p_i = \frac{n_i}{TN}$$

$$\text{If } ((s_i - \mu_1) \leq (s_i - \mu_i))$$

$$\text{Then} \quad p_i = n_i$$

By means of T, we produce binary template on behalf of the following equation:

$$B_{i,j} = \begin{cases} 1 & \text{if intensity of pixel less than T} \\ 0 & \text{Otherwise} \end{cases}$$

The $B_{i,j}$ is finalized ridge in the form of matrix. This result is passing to the next step for getting the result in the form of RGB image. MATLAB code ib2rgb is used for converting binary image to RGB image.

5 Experiments and Results

The below figures demonstrate the investigational outcomes on offered methods and anticipated technique. The proposed method tested using ordinarily image. Using the help of experimental results, we can conclude that this method is robust and effective for image ridge estimation. Figure 4 shows the experimental result on Lenna image.

Table 1 illustrates comparison between offered method and RBCA system using IQP parameters [20]. According to IQP, the MSE should be minimized and PSNR and NOBJ should be maximized as much as possible.

drksujatha23@gmail.com

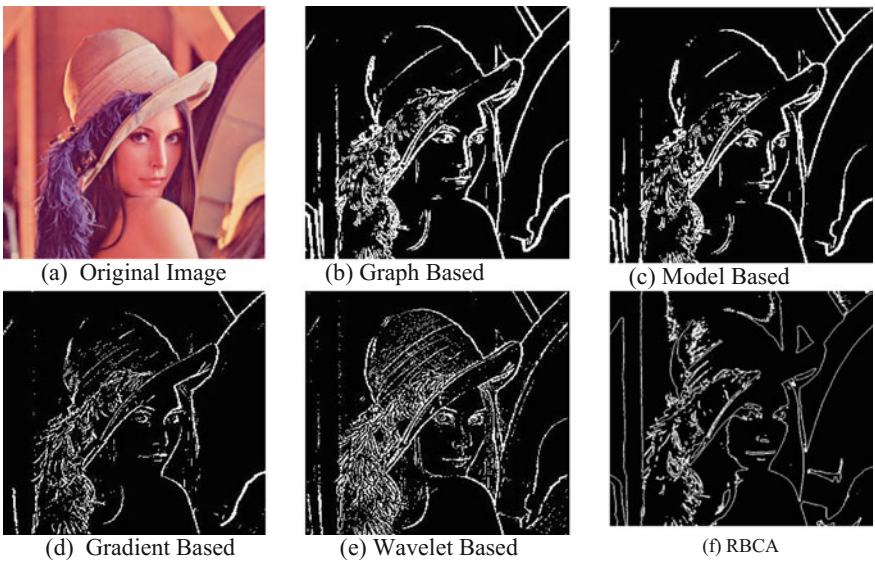


Fig. 4 Experimental results

Table 1 Comparision between existing method and RBCA methods

Quality parameters	Ridge detection methods				
	RBCA	Graph	Model	Gradient	Wavelet
MSE	5.079	5.288	5.536	5.782	5.363
PSNR	5.897	5.698	5.509	5.836	6.072
NOBJ	6.640	5.490	7.710	6.000	5.340

6 Conclusion

The anticipated paper presents sanguine system for image ridge judgment. Different edge judgment and augmentation methods proposed and compared the proposed system with existing methods with the help of IQP parameters. The anticipated technique experimented on various types of images and formed efficacy and heftiness outcomes. Table 1 depicts image quality parameter assessment values of offered methods and RBCA system. On the behalf of IQP values, we state that anticipated RBCA formed improved outcomes. The nearest neighbors up to radius 2 provide improved alternative to calculate a threshold value to achieve planned objective.

drksujatha23@gmail.com

References

1. Pei S-C, Chen L-H (2015) Image quality assessment using human visual DOG model fused with random forest. IEEE Trans Image Process 24(11)
2. Hamamci A, Kucuk N, Karaman K, Engin K, Unal G (2012) Tumor-cut: segmentation of brain tumors on contrast enhanced MR images for radio surgery applications. IEEE Trans Med Imaging 31(3)
3. Ravi R, Josemartin MJ (2014) A novel image processing filter designed using discrete fourier invariant signal. In: 2014 international conference on electronic systems, signal processing and computing technologies. https://doi.org/10.1109/icesc.2014.88. 978-1-4799-2102-7/14 $31.00 © 2014 IEEE
4. Carlisle D, Xu M, Wang J, Yu Z (2012) Image edge enhancement and segmentation via randomized shortest paths. In: 5th international conference on biomedical engineering and informatics (BMEI 2012). 978-1-4673-1184-7/12/$31.00 ©2012 IEEE
5. Yang J, Wright J, Huang TS, Ma Y (2010) Image super-resolution via sparse representation. IEEE Trans Image Process 19(11)
6. Hell B, Kassubeck M, Bauszat P, Eisemann M, Magnor M (2015) An approach toward fast gradient-based image segmentation. IEEE Trans Image Process 24(9)
7. Galbally J, Marcel S, Fierrez J (2014) Image quality assessment for fake biometric detection: application to iris, fingerprint and face recognition. IEEE Trans Image Process 23(2)
8. Yang X, Zhang J, Peng B, You S (2010) An adaptive edge enhancement method based on histogram matching for ultrasound images. In: 2010 international conference on computational and information sciences. https://doi.org/10.1109/iccis.2010.333. 978-0-7695-4270-6/10 $26.00 © 2010 IEEE
9. Cao G, Zhao Y, Ni R, Li X (2014) Contrast enhancement-based forensics in digital images. IEEE Trans Inf Forensics Secur 9(3)
10. Agrawal M, Dash R (2014) Image resolution enhancement using lifting wavelet and stationary wavelet transform. In: 2014 international conference on electronic systems, signal processing and computing technologies. https://doi.org/10.1109/icesc.2014.61. 978-1-4799-2102-7/14 $31.00 © 2014 IEEE
11. Sinha K, Sinha GR (2014) Efficient segmentation methods for tumor detection in MRI images. In: 2014 IEEE conference on electrical, electronics and computer science
12. Zhu Q, Mai J, Shao L (2015) A fast single image haze removal algorithm using color attenuation prior. IEEE Trans Image Process 24(11)
13. Huang J-J, Siu W-C, Liu T-R (2015) Fast image interpolation via random forests. IEEE Trans Image Process 24(10)
14. Duval-Poo MA, Odone F, De Vito E (2015) Edges and corners with shearlets. IEEE Trans Image Process 24(11)
15. Mukherjee A, Kanrar S (2012) Image enhancement with statistical estimation. Int J Multimed Appl (IJMA) 4(2)
16. Zhou M, Geng G (2011) Detail enhancement and noise reduction with true color image edge detection based on wavelet multi-scale. IEEE, pp 1061–1064
17. Senthilkumaran N, Thimmiaraja J (2013) Histogram equalization for image enhancement using MRI brain images. In: 2014 world congress on computing and communication technologies. https://doi.org/10.1109/wccct.2014.45. 978-1-4799-2876-7/13 $31.00 © 2013 IEEE
18. Wang T, Cheng I, Basu A (2009) Fluid vector flow and applications in brain tumor segmentation. IEEE Trans Biomed Eng 56(3)
19. Bartunek JS, Nilsson M, Sällberg B, Claesson I (2013) Adaptive fingerprint image enhancement with emphasis on pre-processing of data. IEEE Trans Image Process 22(2)
20. Yue H, Sun X, Yang J, Wu F (2015) Image denoising by exploring external and internal correlations. IEEE Trans Image Process 24(6)

drksujatha23@gmail.com

Impact of Try A-Gain—An Online Game App for Society



Vijay Singh Rathore, Shikha Maheshwari, Diwanshu Soni, Apoorva Agrawal, Ayush Khandelwal, Aman Vijay, Divyang Bhargava and Aniket Dixit

Abstract In a period where antagonism is found in abundance and suicide rate is expanding at a disturbing rate, such positive applications are precisely what the world needs. Try A-Gain game app is designed to counter those games which are spreading negativity, such as Blue Whale Challenge suicidal game which is known to claim many lives across the globe. The Try A-Gain Game app appears like a lovely method for being appreciative for the life we have and benefit as much as possible from it. This game comes with a set of tasks that will instead have a positive effect on the player's life and helps player to bring him/her closer to self, family, friends, and nation.

Keywords Try A-Gain · Game app · Positivity game app

1 Introduction

The Try A-Gain App is an easily accessible task-based online game, diametrically opposite to the games spreading negativity across the globe such as Blue Whale Challenge game—linked to many deaths [1–3]. The game is developed with an attempt to bring happiness in the users' daily life while encouraging positivity toward life through generous acts using Internet.

It is an incredibly fun game that helps players in living every minute with a positive outlook, staying motivated, being happy, and making others happy. This game app

V. S. Rathore · D. Soni · A. Agrawal · A. Khandelwal · A. Vijay · D. Bhargava · A. Dixit
CSE, Jaipur Engineering College & Research Centre, Jaipur, Rajasthan, India
e-mail: vijaydiamond@gmail.com

A. Khandelwal
e-mail: jindadil.jecrc@gmail.com

S. Maheshwari (✉)
Jaipur Engineering College & Research Centre, Jaipur, Rajasthan, India
e-mail: shikhamaheshwari6583@gmail.com

© Springer Nature Singapore Pte Ltd. 2019

V. S. Rathore et al. (eds.), *Emerging Trends in Expert Applications and Security*,
Advances in Intelligent Systems and Computing 841,
https://doi.org/10.1007/978-981-13-2285-3_26

drksujatha23@gmail.com

comes with a set of tasks that will instead have a positive effect on the player's life and helps player to bring him/her closer to self, family, friends, and nation.

The intention behind developing such game is to help the young minds in a positive way [4]. The game focuses on spreading happy thoughts with its tasks and ends with a message that will bring a smile on anyone's face. The tasks of Try A-Gain app vary from easy workouts like morning walk to socially responsible tasks such as helping parents in household chores, etc. which eventually takes the player closer to the family, friends, and nation and gives reason to live a better life [5, 6].

2 Working of the App

The Try A-Gain app has been uploaded on Google Play Store for download https://play.google.com/store/apps/details.

After launching the game app, if the player is a new user then a signup is required with necessary credentials. Once the signup is completed, the user will get a verification mail from the game app which requires verifying that mail by clicking on the link given in the mail. Visiting user can directly login with user id and password. Also, an option for forgot password is provided in case the user forgets his/her password.

After logging in, the player will see a dashboard having four options, namely, New Challenge, Suggest Task, Edit profile, and Show Submitted Task.

If the user wants to take a new challenge, then he has to choose New Challenge option. Some of the challenges are meeting your old friends, visiting the native places, spending quality time with grandparents, helping mother in her routine kitchen works, taking a test drive at dream car showroom, visiting monuments of historical importance such as Jallianwala Bagh, Sanchi Stupas, Forts, etc., attending family function parents, watching movie with father, etc. Users are required to submit the selfie with completed tasks along with the experience while working on the given tasks. User can perform new challenges only when old ones are verified by other users.

The user can also view his previously completed task by clicking on Show Submitted Task option. In this option, the user can also verify other users' task. Also, there is no option of dislike or disagree with the tasks submitted for verification by the other users to avoid introduction of any negativity with the completed tasks. On the first login, in Edit Profile option, user can change his user id, password, and profile picture. There is also an option that allows user to suggest new tasks.

Till date, there are eight phases with distinct set of tasks. Completion of each phase is marked not only by submitting the tasks but also giving likes and verifying the tasks of other users, which is an attempt to connect users emotionally and is recognized by awarding profile badges to maintain the excitement as well. Users with maximum number of likes will be declared winner of the game.

drksujatha23@gmail.com

2.1 Screenshots of the App

See Figs. 1, 2, 3, 4, 5, and 6.



Fig. 1 Introslider of Try A-Gain



Fig. 2 Sign up page

drksujatha23@gmail.com

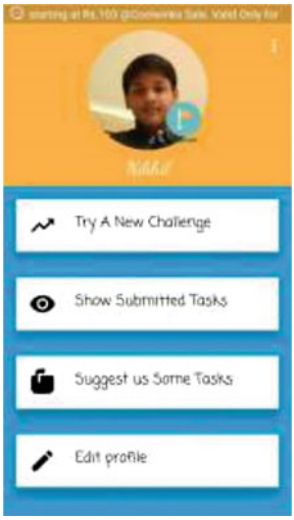


Fig. 3 Dashboard

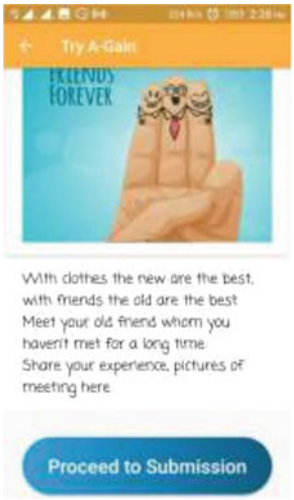


Fig. 4 Problem statement

drksujatha23@gmail.com



Fig. 5 Submitted task



Fig. 6 Badges on successful phase completion

3 Results and Discussion

3.1 Awareness of Try A-Gain Game App

To verify the awareness of Try A-Gain game app against anti-social games such as Blue Whale Challenge [7–9], an area-specific data was collected including approximate 2000 subjects, including players, parents, and students, and the distribution of awareness of both is shown in Table 1 and Fig. 7. More than 4500 subjects never heard of either anti-social/negative games or Try A-Gain game and thus excluded from our further analysis.

drksujatha23@gmail.com

Table 1 Distribution of awareness of Try A-Gain app as compared to anti-social negative games

App name	Player	Parents	Students	Total
Anti-social negative games	6	676	236	918
Try A-Gain game app	478	896	696	2070

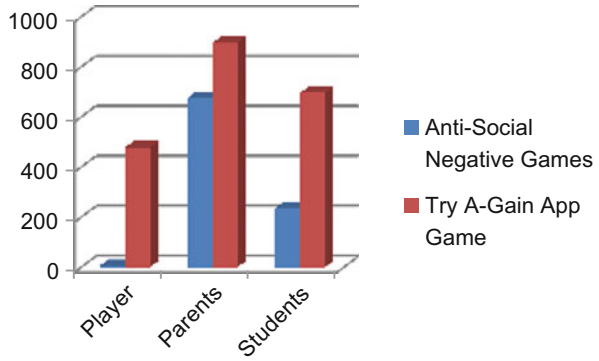


Fig. 7 Awareness comparison in different domains

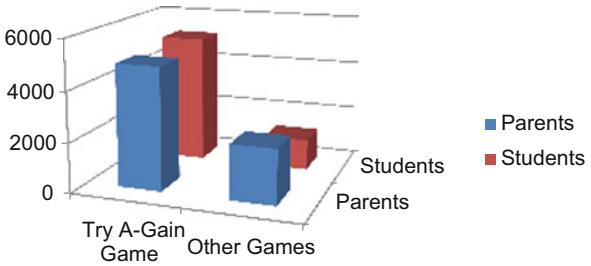


Fig. 8 Comparison of awareness with other games

Figure 8 explores the increment in the awareness related to the other games intended to bring positivity apart from Try A-Gain Game App. As compared to Try A-Gain, other such online game app is less known to the people.

During sampling, a study on identifying the original source of awareness of Try A-Gain Game as well. It was found that a large number of subjects heard about this game through Internet and social media. Awareness programs through newspaper and words of mouth also showed satisfactory results as shown in Fig. 9.

3.2 Change in Attitude and Resultant Impact

To identify the change in behavior and attitude among subjects, the survey was conducted among the undergraduate students of various professional courses. The result is shown using Table 2.

drksujatha23@gmail.com

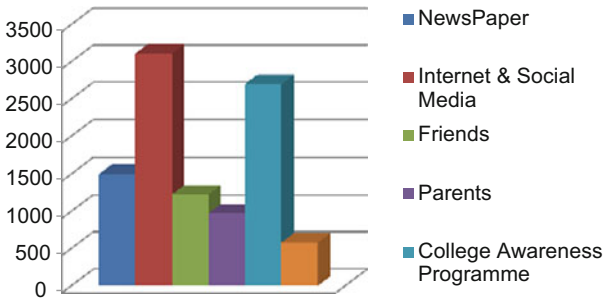


Fig. 9 Sources of awareness

Table 2 Distribution of attitude toward Try A-Gain app

S. no.	Questions (n = 550)	Yes	No
1	Have you tried downloaded the Try A-Gain game?	478	72
2	Were you aware of any other such game app?	214	336
3	Have you tried to download the game app after attending awareness program at college?	109	441
4	Did you find the UI of the game app an interactive one?	421	129
5	Did you find the allotted tasks positive?	478	72
6	Is this app helped you in bringing your childhood memories back?	413	137
7	Do you find tasks of this app associated with self-happiness?	456	94
8	Do you find tasks of this app associated with your family members?	463	87
9	Do you find tasks of this app associated with bringing feeling of nationalism?	397	154
10	Did you enjoy rating of posts of other users?	478	72
11	Did you discuss anything related to task with your parents?	459	91
12	Have you tried promoting this app within your college group?	541	9

4 Conclusion

The Try A-Gain is a game that is spread out by means of online social media and is still growing; it is being propagated by the individuals themselves. On social networks, individuals attempt utilizing all kinds of keywords, hashtags, and images, for capturing the attention of their followers and friends for inviting them to sign up as a player sign up with them.

Great deals of tasks are kept to spread out joy within the good friends and loved ones and therefore propagating the game app to the community. A low portion of players keeps publishing about the tasks completed by them frequently.

Targeted players of Try A-Gain app are either lost contacts or such individuals who are not in touch for a long! Likewise, the interaction in between the users on

drksujatha23@gmail.com

dashboard—that is giving like and complimenting each other's posts—is far more than the interaction in between such users on other social networks.

The appeal of the game app depends on its simplicity to rewind the memories shared on the dashboard. It provides assistance to individuals who may be going through some difficult times or challenge of the life.

5 Limitations and Future Work

API limits of social networks limit the amount of data that can be accessed in the first place. Also, the developers would like to study the characteristics of various users along with their geographical analysis and try to allot them tasks differently to calculate confidence score indicating their level of happiness.

References

1. Panel formed to probe Blue Whale game suicide cases: Government to Delhi HC. Times of India, Oct 2017
2. How to find the Blue Whale game. Higgypop, Aug 2017
3. Russia willing to Assist India in controlling Blue Whale challenge. NDTV, 2017
4. Desk EW (2017) Blue Whale challenge: these are the 5 suspected cases in India. Indian Express, Oct 2017
5. What are the exact 50 challenges in the "Blue Whale challenge"? Reddit, Mar 2017
6. Blue Whale suicide game linked to 130 teen deaths is just tip of the iceberg in the worlds suicide capital Russia. The Sun, 2017
7. Anne (2017) Blue Whale game: fake news" about teens spread internationally, Mar 2017. netfamilynews.org
8. Arora K (2017) Russian social network VKontakte temporarily blocked in India for Blue Whale threat. Times of India, Sept 2017
9. Biswas S (2017) Blue Whale challenge: India ranks no. 1 for highest Blue Whale related searches worldwide, says Google data. India Today, Aug 2017

drksujatha23@gmail.com

A Comparative Analysis of Wavelet Families for Invisible Image Embedding



Neha Solanki, Sarika Khandelwal, Sanjay Gaur and Diwakar Gautam

Abstract In contemporary world, fame of digital video-based applications is among the necessity for copyright protection so as to avoid criminal repeating and distribution of digital data. Digital representation offers many advantages for processing and distributing video and other types of information. Copyright protection adds authentication redundant knowledge in original data like the possession details and owner logo within the digital media while not compromising its sensory activity quality. In this article, we analyze the performance of wavelet functions, viz., Haar, Daubechies, symmetric, and biorthogonal for invisible image embedding. The embedding and extraction processes are deployed using the number theory concept of embedding.

Keywords Image watermarking · Image segmentation · Wavelet function Entropy

Neha Solanki created the first draft of this document.

N. Solanki (✉) · S. Khandelwal
Geetanjali College of Technical Studies, Udaipur, India
e-mail: nikisolanki22@gmail.com

S. Khandelwal
e-mail: sarikakhandelwal@gmail.com

S. Gaur
Jaipur Engineering College & Research Centre, Jaipur, India
e-mail: sanjay.since@gmail.com

D. Gautam
Sharda University, Greater Noida, UP, India
e-mail: diwakar.gautam1@sharda.ac.in

© Springer Nature Singapore Pte Ltd. 2019

V. S. Rathore et al. (eds.), *Emerging Trends in Expert Applications and Security*,
Advances in Intelligent Systems and Computing 841,
https://doi.org/10.1007/978-981-13-2285-3_27

drksujatha23@gmail.com

1 Introduction

Digital representation offers many facilities for processing and distributing video and different kinds of data. First, digital software packages offer unparalleled developing, enhancing, presenting, and flexibility in manipulating digital facts. Analog gadgets lack the power, malleability, and extensibility of software program processing. On some of these structures, existing open and proprietary protocols which include the World Wide Web permit any user to speedy and inexpensively acquire, provide, alternate, and find digital statistics. Lastly, virtual facts can be processed, and particularly, copied without introducing loss, degradation, or noise. For instance, an infinite number of ideal copies may be constituted of a single digital video signal. In assessment, the addition of noise into a duplicate from analog sign processing is unavoidable.

2 Literature Survey

Video watermarking formed is likewise classified into two significant guidelines focused on the approach of undertaking watermark bits inside the host video. The two classes are spatial and rebuild the area. The spatial zone is watermarking where inserting and location of the watermark are done by the method for hetero controlling the stage profundity estimations of the video body. Change area techniques exchange reflection estimations of the host video reliable with a pre-decided model and gadgets additional viable than deliberation put contraption. The foremost techniques deployed Discrete Cosine Transform (DCT) and the Discrete Wavelet Transform (DWT) [1]. Video watermarking alludes to inserting watermarks in a total video from prohibited redundancy and distinguishes controls. A spread of tough and delicate video watermarking methods is pondered to determine the restricted redundancy and confirmation of ownership issues in any case on seeing manipulations [2]. The systems are frequently partitioned into methodologies that work of art on packed or uncompressed information. More than a couple of sorts of watermarking plans are anticipated for differing programs. For copyright-related capacities, the inserted watermark is foreseen to be resistant to various types of noxious and non-malevolent controls to some degree, as long in light of the fact that the controlled substance material stays to be profitable as far as industry significance or important regarding errand uncommon. Thus, watermarking plans for copyright-related applications are on the whole intense [3]. Computerized watermarking practically incorporates implanting a key stylish mystery motion into advanced gifts in a strong and undetectable approach. Moreover, this basic flag is

drksujatha23@gmail.com

deliberately attached to the host aptitudes all together that it survives advanced to simple change. There is a luxurious change-off among three parameters: capacities payload, steadiness, and general execution. The essential vitality offered by overhaul territory systems is that they may pick up from exceptional living arrangements of exchange areas to deal with the confinements of pixel-based techniques [4]. While there are various solid watermarks in the DCT zone, there are relatively less existing records of movement in watermarking systems in DCT space [5]. Kim [6] entered watermark bits as pseudo-arbitrary successions inside the recurrence area. Langelaar [7] camouflaged watermarks with the guise of wiping out or holding select DCT coefficients. Borg [8] disguised the watermark in JPEG photographs by utilizing driving pick DCT squares to fulfill positive direct or roundabout imperative. Some inserts watermark designs inside the quantization module, while DCT [9] or specifically pieces upheld human visual models. Choi [10] used square connection by driving DCT coefficients of a piece to be bigger or littler than the average of the neighbor squares. In 1995, Cox advanced a fresh out of the box new arrangement of tenets of the utilization of spread range to insert a stamp [11].

3 Image Segmentation Based on Entropy Measures

The object segmentation in this article is conducted using entropy-based image segmentation technique. The methodology of uncertainty-based image segmentation relies on the use of the grey co-occurrence matrix C(m1, m2) and Shannon uncertainty measure. In this article, we opted for Shannon entropy measures against Renyi, Havrda-Charvat, Kapur, and Vajda entropy on image functions [11]. The primary steps of the algorithm are presented here for the sake of comfort:

i. Evaluate the intensity value co-occurrence matrix $C(m_1, m_2)$, for each neighborhood orientation and cumulate it all.
ii. Evaluate two-dimensional probability distributions $C(m_1, m_2)$/(Number of pixels in the given image).
iii. Using prob. matrix, find the entropy value related to every possible threshold *t*.

Find the minimal regional minimas, inferring the informative points. Among all regional minima, select the minimum regional minima.

iv. The *t* value of the minimum regional minima is the optimum threshold for image segmentation.

drksujatha23@gmail.com

3.1 Entropy Measures

I. Shannon Entropy:

Shannon's entropy measure provides an absolute limit on the best possible lossless compression of a signal under certain constraints [12]. It is defined as

$$H_s\left(p_{m_1,m_2}\right) = -\sum_{m_1}\sum_{m_2} p_{m_1,m_2}\log p_{m_1,m_2}$$

where 2-D random variables are associated with the unified probability distribution. In this major project work, we have computed the segmentation threshold evaluated from the entries of the gray-level co-occurrence matrix [5, 6] of the given image as given by the relation:

$$p_{m_1,m_2} = C_{m_1,m_2}/(MN)$$

4 Embedding Strategy and Experimental Outcomes

The experimentation performed over MATLAB2013a running on Intel Core I5 processor operating at 5.5 GHz. Windows 7 is the basic platform to execute MATLAB commands from higher level to lower level. The following set of algorithms is tested for embedding and extraction of embedded watermark.

Algorithm 1

I Embedding Procedure

Step1. Apply single-level two-dimensional wavelet transformation on the input image.
Step2. Subdivide the resulted LH and HL band into noninteracting subblocks of size 2 × 2. In the next stage, the message bit is embedded in odd columns of LH band and in even columns of HL band, respectively.
Step3. Given each selected block $B(m, n)$ and message bit w, evaluate the mean value for each block of $B(m, n)$, as depicted below:

$$M(m,n) = \frac{\sum_{i=0}^{1}\sum_{j=0}^{1}\left(x_{m+i,n+j}\right)}{4}$$

drksujatha23@gmail.com

Now embed the bit *w* in the host image using below steps:

- R := Modulus of six resulted from *M(m,n)*;
- **for** i := 0 **to** 1
- **for** j := 0 **to** 1
  - **if** $0 \leq R < 3$ **then**
    - **if** w = 1 **then** $x_{m+i,n+j} := x_{m+i,n+j} + (3-R)$;
    - **if** w = 0 **then** $x_{m+i,n+j} := x_{m+i,n+j} - R$;
  - **if** $3 \leq R < 6$ **then**
    - **if** w = 1 **then** $x_{m+i,n+j} := x_{m+i,n+j} + (3-R)$;
    - **if** w = 0 **then** $x_{m+i,n+j} := x_{m+i,n+j} + (6-R)$;

Step4. Retransform back from wavelet to spatial domain to get the watermarked image.

II Extraction Procedure (To extract the hidden message from watermarked image)

Step1. Apply single-level two-dimensional wavelet transformation on the watermarked image.
Step2. Subdivide the resulted LH and HL band into noninteracting subblocks of size 2×2. In the next stage, the message bit is extracted from odd columns of LH band and from even columns of HL band, respectively.
Step3. Given each selected block $B(m, n)$ and message bit w, evaluate the mean value for each block of $B(m, n)$, as depicted below:

$$M(m,n) = \frac{\sum_{i=0}^{1} \sum_{j=0}^{1} \left(x_{m+i,n+j}\right)}{4}$$

Now extract the bit *w* from watermarked image using below steps:

- R := Modulus of six resulted from *M(m,n)*;
  - **if** $0 \leq R < 1.5$ **then** w:= 0;
  - **if** $1.5 \leq R < 4.5$ **then** w:= 1;
  - **if** $4.5 \leq R < 6$ **then** w:= 0;

drksujatha23@gmail.com

5 Results and Parameters

Computerized watermarking innovation is a propelling field of registering, cryptology, flag process, and interchanges. The watermarking investigation is extra energizing as it wants to aggregate thoughts from every field alongside human psychovisual examination, sight and sound framework, and workstation graphics. The watermark may be noticeable or undetectable sort; each independently has its applications. The test is led on a 4Ghz I5 processor, with the examined calculation mimicked in MATLAB-2013 A running on Windows platform. Figure 1a, b represents the cover images and watermark images set.

The segmentation result for the message images using Otsu method is depicted as below in the left column of Fig. 2.

The watermarked images resulted from embedding algorithm applied individually are depicted as above in right column of Fig. 2.

Here, binary message bits are embedded into LH and HL bands of cover images. Extracted watermarks at user end using Algorithm 2 is depicted in Fig. 3. The PSNR and NC values calculated using the watermarked images of Fig. 1 and extracted watermarks of Fig. 3 for various wavelet functions are tabulated as below (Table 1).



Fig. 1 Cover and message image set

(a) Cover Images **(b) Message Image**

drksujatha23@gmail.com



Fig. 2 The watermarked cover images

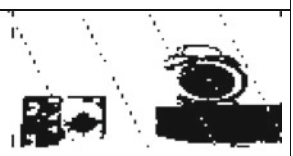
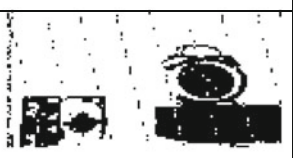


Fig. 3 The extracted watermark or message at user end

It can be inferred from the above table that PSNR and NC coefficients are playing a major role in comparing various wavelet functions for image embedding and extraction. The highest value of PSNR is attained using Daubechies wavelet. It due to fact that Daubechies function is more better realization for decomposition and reconstruction of signal as compared to any other wavelet function.

drksujatha23@gmail.com

Table 1 Peak signal-to-noise ratio (PSNR) and normalized coefficient (NC)

Wavelet function	PSNR	NC	Watermaked image	Extracted message
Haar	105.67	0.9821		
Daubecheis	130.54	0.94		
Symlets	120.21	0.95		
Biorthogonal	110.86	0.96		

6 Conclusion

As a future scope, the thought of cryptography and digital watermarking may be combined to implement safer digital watermarking system. In this article, we analyzed the performance of proposed encryption model deploying various wavelet functions for information hiding. Haar, Daubechies, Symlets, and biorthogonal wavelet functions are incorporated to decompose a given image in wavelet coefficients, and the given message image is embedded into alternate rows and columns of LH and HL band, respectively.

The future aspect of related work is listed as below:

i. Application of other segmentation schemes can be tested for improved performance either in terms of PSNR, NC, or computational time.

drksujatha23@gmail.com

ii. The suggested embedding scheme can be extended to video watermarking, where watermarked frames will be an add-on secrecy point.

Proposed scheme can be processed through code conversion procedures for its hardware-level implementation for video processors.

References

1. Yeo, Yeung MM (1997) Analysis and synthesis for new digital video application. In: International conference on image processing (ICIP97), vol 1, p 1, 1997
2. Natarajan M, Makhdumi1 G (2009) Safeguarding the digital contents: digital watermarking. DESIDOC J Libr Inf Technol 29:29–35
3. Podilchuk CI, Delp EJ (2001) Digital watermarking: algorithms and applications. Sig Process Mag 18:33–46. IEEE
4. Doerr G, Dugelay JL (2004) Security pitfalls of frame-by-frame approaches to video watermarking. IEEE Trans Sig Process 52:2955–2964
5. Thakur MK, Saxena V, Gupta JP (2010) A performance analysis of objective video quality metrics for digital video watermarking. In: 3rd IEEE international conference on computer science and information technology (ICCSIT), 2010, vol 4, pp 12–17
6. Voloshynovskiy S, Pereira S, Pun T (1999) Watermark attacks. In: Proceedings of Erlangen watermarking workshop 99, Oct 1999
7. Langelaar G, Setyawan I, Lagendijk R (2000) Watermarking digital image and video data: a state of art overview. IEEE Sig Process Mag 20–46
8. Hartung F, Girod B (1998) Watermarking of uncompressed and compressed video. Sig Process 66(3):283–301
9. Lee J et al (2001) A survey of watermarking techniques applied to multimedia. In: IEEE International Symposium on Industrial Electronics, vol 1, pp 272-277, 2001
10. Cox et al (2002) Digital watermarking: principal and practice. Morgan Kaufmann
11. Meng J, Chang S (1998) Embedding visible video watermarks in the compressed domain. In: Proceedings of international conference on image processing, ICIP 98, vol 1, pp 474–477, 1998
12. Meggs PB (1998) A history of graphic design, 3rd edn. Wiley, pp 58. ISBN 978-0-471-29198-5

drksujatha23@gmail.com

A Video Database for Intelligent Video Authentication



Priya Gupta, Ankur Raj, Shashikant Singh and Seema Yadav

Abstract In this paper, we depict a special video database which comprises the genuine snapshots of individuals and items, caught under different light conditions and camera positions. We have arranged every one of the recordings of our database into six classifications, out of which four classifications depend on the developments of camera and articles (caught by the camera). The rest of the classes of the database are sunshine recordings and night vision recordings. The recordings caught under the regular light source (for example, daylight) are canvassed in sunlight recordings class. The night vision recordings class has an indistinguishable setup and condition from the sunshine recordings classification; however, the recordings are caught in low light condition and the camera is recording in night vision mode. Every class of this video database offers a decent circumstance for the test of video validation and to understand the believability of video confirmation calculations as well. We have connected our own particular clever video validation calculation on every classification of the video database and get the outcomes with the general precision of 94.85%, subjected to different altering assaults.

Keywords Development · Reconnaissance · Abundant

P. Gupta · A. Raj (✉) · S. Singh · S. Yadav
JECRC, Jaipur, Rajasthan, India
e-mail: ankurraj.cse@jecrc.ac.in

P. Gupta
e-mail: priyagupta.cse@jecrc.ac.in

S. Singh
e-mail: shashikant.cse@jecrc.ac.in

S. Yadav
e-mail: seemayadav.cse@jecrc.ac.in

© Springer Nature Singapore Pte Ltd. 2019

V. S. Rathore et al. (eds.), *Emerging Trends in Expert Applications and Security*,
Advances in Intelligent Systems and Computing 841,
https://doi.org/10.1007/978-981-13-2285-3_28

drksujatha23@gmail.com

1 Introduction

The utilization of the video reconnaissance is multiplied in different parts of our everyday lives. A normal American resident crosses around 500 CCTV cameras in a day. The exploration on video-based observation frameworks opens the entryway of different testing issues to the scientists, for example, video confirmation, video-based human ID, and check (video biometrics) for security frameworks and protest following [1, 2]. The examination fundamentally depends on the recordings, taken under generally controlled perspective and light conditions. In view of different requirements in the previous decade, advance has been made to tackle the complex issues [3–5]. In video confirmation, existing strategies confront another sort of issue each day. Since the watermarking and advanced mark-based video confirmation systems require pre-details of equipment, they do not get the job done; the honest to goodness brings about official courtroom for foul recordings. Scientific specialists utilize factual devices and systems for the confirmation of such sorts of recordings. Crude recordings have an extraordinary level of assorted variety. Promote different altering assaults [6] and increment the multifaceted nature of the issue of video confirmation. However, clever verification strategy [7, 8] gives momentous outcomes, and they as a rule take a shot at the recordings which catches the regular development of individuals and different from various edges. The recordings taken in these enlightenment conditions are the principle issue for confirmation purposes, and this would turn into a critical range of future research. One of the critical explanations behind the moderately constrained measure of research on the verification of crude recordings, taken in characteristic or manufactured enlightenment conditions, is the absence of a standard video database, which gives countless in an assorted variety of settings and imaging circumstances [9]. There is an abundant need of such sort of video database to test the precision and effectiveness of the validation systems for crude recordings in which the items (counting individuals) and camera have the relative developments and in which the brightening comes either from a characteristic light source or from counterfeit light sources. Notwithstanding confirmation-based applications, there is another utilization of video database, containing characteristic developments of individuals and articles, for human recognizable proof and check in video biometrics [10]. This research paper depicts a database of crude recordings of various questions and individuals. This database is fundamentally created for testing the precision and effectiveness of wise video validation strategies against different altering assaults [6, 7]. Be that as it may, other validation procedures (particularly custom fitted for some altering assaults) can likewise be tried over this video database.

2 Database Definition

We have made a video database of 120 recordings, initially recorded by a Mini DV (HCR DC 38) SONY Handy cam in different enlightenment conditions and camera positions. A portion of the recordings of the database was taken at shut in

drksujatha23@gmail.com

appropriate light, and others are taken under regular light (daylight) condition in open-air situations. The 80 recordings, out of 120 recordings of the database, are recorded in different places of camera and items (caught by the camera) and ordered into four classifications which are moving camera moving articles, moving camera still protests, still camera moving items, and still camera still questions. Here, every class has 20 video cuts. The staying 40 recordings are recorded in various light conditions and ordered into two classes. These are light recordings and night vision recordings, each containing 20 video cuts. The sunshine recordings and night vision recordings of the database were set up in two sessions of the chronicle. To start with, first session, which incorporates the sunlight recordings, finishes under regular enlightenment condition. The second session, which incorporates the night vision recordings, comprises same setup as was in sunlight recordings and finishes under low light condition. The normal interim among first and second sessions was 6 h. After the planning of this unique video database, we have stretched out our database as per different altering assaults [6]. Additional précising recording subtle elements, for example, camera separate, and so forth, are given in the appendix. These classifications are portrayed in a nutshell in the subsections beneath.

2.1 Moving Camera Moving Articles (MCMO)

In the MCMO classification, the camera is moving either along a bearing or along a pivot amid the account. Here, the development of camera is not extremely quick, yet at a fix rate. What is more, in the scene, which is being caught by the camera, objects are additionally moving. Here, the articles developments can fluctuate.

2.2 Moving Camera Still Objects (MCSO)

In the MCSO class, the camera is again moving along any heading or a hub; yet, the articles, being caught by the camera, are not moving. Still protests, for example, divider, painting, statues, and landmarks, are caught in the recordings of this class.

2.3 Still Camera Moving Object (SCMO)

The SCMO classification settles the camera at a point and catches the recordings. There is no development of camera in this class; however, the articles, which are being caught by the camera, have the developments.

drksujatha23@gmail.com

2.4 Still Camera Still Object (SCSO)

As the name recommends, this classification has the recordings of settled articles (both foundation and closer view items) and they are shot by a settled camera.

2.5 Daylight Videos

The videos shot in the daylight condition are put in this category. Here, we fixed the camera at a certain position and shoot the video in natural illumination condition. Most of the videos of the category of daylight videos are captured in outdoor environments.

2.6 Night Vision Videos

The recordings of fundamental class have shot in the night (low light condition). For this classification, we have made an indistinguishable setup from which was in the sunlight condition, i.e., we settled the camera at a similar area and position (as in light condition) and shot the video in the night. Here, the camera is recorded in night vision mode.

The recordings, shot in the sunshine and night vision conditions, are recorded such that the half segment of the classifications (light and night vision) contains the sunlight and night vision recordings with various items (closer view objects) in the scene and half part of the classes contains the sunlight and night vision recordings with similar questions in the scene. We have expanded this database and tried a wise video validation calculation, proposed in [2], over this broadened video database against different altering assaults.

3 Extension of the Video Database

The intention behind the augmentation of this video database was to test the precision of any smart video verification calculation against different altering assaults. We have broadened our database such that the entire database contains two sections. One player in the database comprises non-altered recordings of the considerable number of classifications and another part comprises altered recordings of the considerable number of classes, subjected to different altering assaults [6]. Since the smart (learning based) video confirmation calculation (proposed in [1]) utilizes test information (recordings) for the preparation reason, we have part each initially recorded recording of the considerable number of classifications of the video database into ten pieces, all having measured up to number of casings. The purpose

drksujatha23@gmail.com

for this part is that a video may have distinctive bits of occasions in various casing groupings. Consequently, we have added up to 1200 non-altered recordings. The more exact points of interest of the database are as per the following:

Our database contains 120 really recorded non-altered recordings with 5040 casings each. Each caught at 24 fps. This video information is utilized as the basic truth. These 120 recordings are characterized into 6 classes; each containing 20 recordings and every video of the considerable number of classifications are additionally part into 10 squares with estimated recordings of 504 edges, as said above. For each of the 20 principal truth recordings of each classification of the video database, distinctive altered duplicates are made by subjecting them to various video altering assaults [6]. The points of interest of the formation of other (Tampered) some portion of the initially recorded video database are given underneath:

- For every video, 10 duplicates are made with outline evacuation assault in which 1 to 20 edges of the video have been dropped aimlessly positions.
- For outline expansion assault, we initially select a video inside the class. Casings of this video are embedded indiscriminately positions in the rest of the recordings of that specific classification of the video database to create 10 altered duplicates of every video of that specific classification, barring the chose video. This operation is performed for every class of the video database to produce altered duplicates of all ground truth recordings.
- For spatial altering assaults, we utilized proficient programming. With the utilization of this product, we can modify the topic of the casings. This change is performed in different perspectives, for example, protest expansion and question expulsion from the casings. Ten duplicates of every video of the considerable number of classifications of video database are made, subjected to spatial altering assaults.

We, in this way, have added up to 120 ground truth recordings with 1200 non-altered recordings, 1200 recordings with outline option assault, 1200 recordings with outline expulsion assault, and 1200 recordings with spatial altering assaults.

4 Application of an Intelligent Authentication Algorithm Over the Video Database

We have connected our clever video confirmation calculation, proposed in, over this video database. The calculation utilizes support vector machine, which is a non-direct classifier. Bolster vector machine is a capable approach for taking care of issues in nonlinear grouping, work estimation, and thickness estimation [2]. It makes an association among edge and furthermore can compute the measurable nearby data in the casings. A Support Vector Machine (SVM) [10]-based learning

drksujatha23@gmail.com

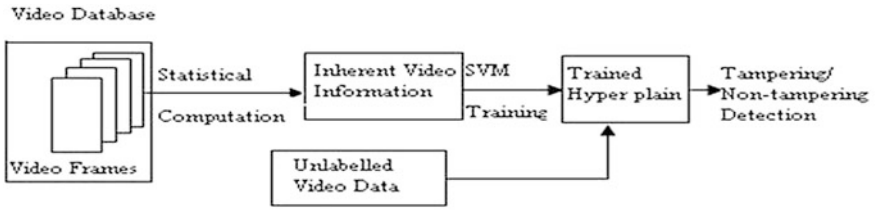


Fig. 1 Block diagram of intelligent video authentication algorithm

calculation is then used to group the video as altered or non-altered. The idea of the calculation is appeared in Fig. 1. It covers the two sorts of altering assaults, spatial and transient. As a result of SVM-based preparing, it can tell the contrast amongst assault and worthy operations. The means associated with the preparation of the bit and alter recognition with order are clarified beneath.

In every class of the video database, non-altered recordings have just about a set example of changes in back-to-back edges of the recordings. The camera catches the edges of the video at a settled rate, and the development of items and camera extends in a short space. While on account of altered recordings, a considerable measure of assortments is there. We thought with a mentality of a malevolent aggressor when we arranged the altered duplicates of our ground truth recordings. That is the reason we have utilized here altered recordings a large portion of the circumstances, for the approval procedure of our calculation (Figs. 2 and 3).

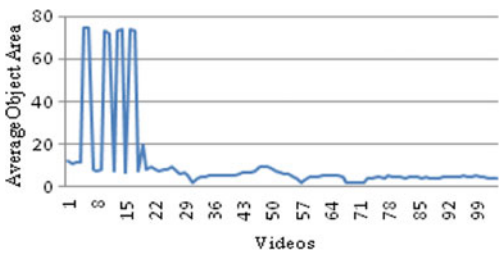


Fig. 2 Plot of average objects area as statistical local information

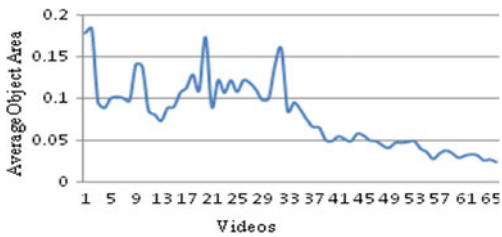


Fig. 3 Plot of object areas where videos are spatial tempering attack

drksujatha23@gmail.com

Table 1 Classification results of the video authentication algorithm for each category of the video database

Database categories	Total number of videos	Number of correctly classified videos	Classification accuracy (%)
MCMO	303	290	95.71
MCSO	233	213	91.41
SCMO	328	310	94.51
SCSO	159	159	100
Daylight videos	265	247	93.21
Night vision videos	267	256	95.88
Total	1555	1475	94.85

Our calculation gives the best outcomes for SCSO classification, in which everything is settled at its position. For all the three altering assaults, in the SCSO class of our video database, the calculation plays out the order with the 100% precision. In any case, our calculation performs well likewise for whatever remains of the classes of our video database.

$$K(x, x_i) = exp\left(-\gamma \parallel x - x_i \parallel^2\right), \quad \gamma > 0$$

where X and Xi represent the input vectors and v is the RBF parameter [9].

The outcomes given in Table 1 abridge the execution of the video verification calculation. With the general arrangement precision of 94.85%, our calculation gives great outcomes for every class of our video database exclusively.

5 Summary and Future Scope

This video database is essentially outlined and arranged for concentrate the execution of video validation calculations, particularly smart video verification calculations. Here, each classification of the database is a testing situation, in which the legitimacy of the recordings should be set up. We have connected a savvy video validation strategy, proposed in [10], on every class of the video database and get the outcomes with general grouping exactness of 94.85%. In future, we might want to extend our video database, which would cover some more basic conditions for video recording, similar to the recordings caught in a circumstance where the camera and items are moving quick, recordings in basic climate conditions, and the recordings caught in unsafe condition, and apply the insightful verification calculations for getting the outcomes with respect to a wide range of altering assaults.

drksujatha23@gmail.com

Appendix

A. Equipment

The recordings of the database have been gathered utilizing SONY DCR-HC 38 MiniDV Handycam camcorder with 40X optical zoom. DCR-HC 38 utilizes 1/6-inch progressed HAD (Hole Accumulation Diode) CCD imager with 340 k pixels which thus gives great lucidity video (up to 500 lines of flat determination). It additionally has the highlights like super steady shot picture adjustment framework which utilizes movement sensors to recognize and make up for the shaking of camera without trading off picture quality, and the night shot plus infrared framework, which records the subjects up to 10 feet away utilizing the implicit infrared framework. We have recorded the recordings in two distinct situations: indoor and open-air conditions.

B. Outdoor Environment

Just the half of the night vision classification recordings were recorded in fake light condition, in open-air condition. The rest classifications of the database have the recordings, recorded in common light condition, in open-air condition. In this condition, we have brought the recordings with a direct range in which the separation between the camera and items ranges from 3 to 13 m.

C. Indoor Environment

The indoor condition has the half of the night vision classification recordings, recorded in manufactured and lowlight condition. For the vast majority of the recordings, indoor condition has a fake light setup. In this condition, we have brought the recordings with a short proximity in which the separation among camera and articles ranges from 1 to 3 m. Each class has a few recordings, recorded in indoor condition.

References

1. Guerrini F, Leonardi R, Migliorati P (2004) A new video authentication template based on bubble random sampling. In: Proceedings of the European signal processing conference
2. Yan Q, Kankanhalli MS (2003) Motion trajectory based video authentication. In: Proceedings of ISCAS, vol 3, pp 810–813
3. Ulrich P, Wollinger GR (2011) A surveillance studies perspective on protest policing: the case of video surveillance of demonstration in Germany. J Soc Mov 3(1):12–38
4. Hsia S-C, Hsiao CH, Huang C-Y (2009) Single-object-based segmentation and coding technique for video surveillance system. J Electron Imaging 18(03):033007
5. Devasena CL, Revathi R, Hemalatha M (2011) Video surveillance systems-a survey. IJCSI Int J Comput Sci Issues 8(4). No 1. ISSN (Online) 1694-0814
6. Upadhyay S, Singh SK (2012) Video authentication: issues and challenges. IJCSI Int J Comput Sci Issues 9(1). No 3. ISSN (Online) 1694-0814

drksujatha23@gmail.com

7. Yin P, Yu HH (2012) Classification of video tampering methods and countermeasures using digital watermarking. In: Proceedings of SPIE, multimedia systems and applications IV, vol 4518, pp 239–246
8. Singh R et al (2008) Integrating SVM classification with SVD watermarking for intelligent video authentication. Telecommun Syst J. Special issue on computational intelligence in multimedia computing. Springer
9. O'Toole AJ et al (2005) A video database of moving faces and people. IEEE Trans Pattern Anal Mach Intell 27(5)
10. Vapnik VN (1995) The nature of statistical learning theory. Springer

drksujatha23@gmail.com

Software Quality Improvement Through Penetration Testing



Subhash Chandra Jat, C. S. Lamba and Vijay Singh Rathore

Abstract In this paper, we explore the use of service-level agreements to improve the quality and management of software-intensive systems. Software quality is typically used in outsourcing contracts for post-production support. We propose that software quality be used in software acquisition to support quality and process control throughout the lifecycle of a software-intensive system. The hypothesis was tested using two methodologies. The principal strategy clarified how software quality could be utilized all through a framework's lifecycle to enhance programming quality. Programming quality could be utilized to enhance general quality in the improvement exertion and at last item. Real concentration of this work is to create real programming quality for a particular lifecycle stage to represent the ideas of programming quality and to show their incentive as a quality control and testing apparatus. Programming improvement is the way toward coding usefulness to meet characterized end client needs. While programming testing has a tendency to be viewed as a piece of improvement, it is truly its own teach and ought to be followed as its own venture. Programming testing, while working intimately with advancement, ought to be sufficiently free to have the capacity to hold-up or moderate item conveyance if quality destinations are not met. In this work, we concentrate on the penetration testing way to deal with enhancing the product quality. Infiltration testing is utilized to look for vulnerabilities that may show in a system framework. The testing procedure for the most part includes reenacting distinctive sorts of assaults on the objective of a machine or system. This sort of testing gives a composed and controlled approach to recognize security issues. For the most part, the assets and time required for far-reaching testing can make penetration testing cost concentrated. Therefore, such tests are generally just performed amid critical

S. C. Jat (✉)
Department of Computer Science, RCEW, Jaipur, India
e-mail: subhashccjat@yahoo.com

C. S. Lamba · V. S. Rathore
Jaipur Engineering College & Research Centre, Jaipur, Rajasthan, India
e-mail: professorlamba@gmail.com

V. S. Rathore
e-mail: vijaydiamond@gmail.com

© Springer Nature Singapore Pte Ltd. 2019

V. S. Rathore et al. (eds.), *Emerging Trends in Expert Applications and Security*,
Advances in Intelligent Systems and Computing 841,
https://doi.org/10.1007/978-981-13-2285-3_29

drksujatha23@gmail.com

turning points. An infiltration test is a strategy for assessing the security of a PC framework or system by reproducing an assault from pernicious pariahs as well as insiders. A few techniques did amid infiltration tests can be effortlessly automatized.

Keywords Software quality · Penetration test · Testing environment

1 Introduction

The fast development in the web and web innovations has been advantageous to organizations and people groups. With the ascent of new advances comes the test of giving a protected domain to the productive preparing. A review directed by the CISCO in 2013 recommends that more than 90% of IT-based organizations have succumbed to pernicious assaults [1]. Security testing is utilized to fabricate a protected framework; however, it has been overlooked for quite a while. It is of impeccable significance nowadays for all the IT security people groups. In this day and age, protection and security have been relegated premier significance, and along these lines, it is exceptionally prescribed to look forward for information and operations' security in programming applications, which requests earnest consideration, yet it is fairly disregarded. Hence, our goal is to present engineers with a regarded significance of framework's security, which can be actuated by executing security testing philosophy in SDLC procedure to deliver a safe programming framework. Thus, security testing has been characterized from designer's perspective. It looks like strategies that should be acquired in SDLC procedure to join security include in programming. Programming Security Unified Knowledge Architecture depicts security testing's qualities and targets as well as gives some designer's rules to deliver a safe programming framework.

Prior to an infiltration test, certain key issues should be submitted in request to guarantee valuable and convenient outcomes. It incorporates the specialized prerequisites, for example, time requirements; cover the full scope of the dangers, the scope of IP addresses over which the test is to be led, and the frameworks that are to be assaulted and furthermore those that are not to be assaulted as a component of the test with negligible disturbance to typical operation. Different prerequisites may likewise incorporate lawful and legally binding issues determining risk data to people with respect to the test occurred. Such prerequisites can fluctuate contingent upon legitimate structures in the association or even the host nation of the association.

Arrange infiltration testing is a notable approach utilized for security testing. Infiltration testing can be a relentless errand which depends much on human learning and ability, with different strategies utilized, and a broad measure of apparatuses utilized as a part of the procedure. A deliberate way to deal with penetration testing is in this manner prescribed. The blemish speculation approach, utilized as a part of this postulation, speak to a standout among the most utilized

drksujatha23@gmail.com

models for infiltration testing and have incredible likenesses in other penetration testing procedures and guidelines utilized today.

There are few purposes behind an association to contract a security expert to play out an entrance test. The primary reason is that security breaks can be to a great degree expensive. An effective assault may prompt to coordinate money-related misfortunes, hurt the association's notoriety, trigger fines, and so on. With an appropriate infiltration test, it is conceivable to distinguish security vulnerabilities and after that take countermeasures before a genuine assault happens. An entrance test is by and large performed by individuals outer to the association in charge of the framework under test. Subsequently, the analyzers work with an alternate perspective of the framework's assets and might have the capacity to distinguish issues that were not promptly noticeable to inside administrators.

2 Literature Review

We have reviewed some earlier efforts to automate the penetration testing process. There are various tools which provide the basis for understanding the automated procedures for penetration testing in the context of their production environments. A business application created for automatized penetration testing created by Core Security Technologies. Center security's impact is GUI-based application intended for facilitating the work of corporate security instrument which needs an effective application to perform infiltration testing on their frameworks [2]. This application computerizes all periods of an infiltration test, from necessity detail to definite report era. Fundamental idea driving this application is system utilized by the dominant part of automatized infiltration testing instruments, for example, the begin checks a scope of hosts in a system, searching for vulnerabilities for which it has reasonable endeavors. In an extra way after the powerlessness misuse, this application can introduce operators on the influenced machines that give distinctive levels of remote get to. These dynamic operators can dispatch extra tests from the new area, permitting the infiltration analyzer to move from host to have inside the framework under test. The adventures utilized by this product are continually overhauled and accessible to the end clients. The accessible endeavor database contains a substantial number of a la mode abuses which gives it the capacity to test an extensive variety of frameworks. Real downside of Core Security Technologies programming is its high cost and the absence of a summon line interface.

Another business application created for automatized penetration testing created by Immunity Inc [3]. Resistance's Canvas is a weakness misuse apparatus that utilizes an indistinguishable approach from Core Impacts, the main distinction; it gives a lower level of computerization and it has less components, for example, turning and automatized revealing. Significant points of interest of this device over Core Impact are an extensively bring down cost and an element of order line interface. With respect to extra point, this application does not give completely computerized techniques to infiltration testing. It is a fundamental bolster device for

drksujatha23@gmail.com

infiltration analyzers; those can utilize it to assemble data about the framework under test and pick fitting endeavors for activities among all gave. This apparatus can computerize parts of the penetration testing process; the end client of this instrument must have a considerable information about infiltration testing and framework security.

Quick Track [4] is a python-construct open-source extend situated in light of the Metasploit structure giving entrance analyzers automatized instruments to recognize and misuse vulnerabilities in a system. Quick Track develops Metasploit with extra elements and is made out of a few devices worried with various parts of the infiltration test: MSSQL server assaults, SQL infusion, Metasploit Autopwn Automation, Mass Client-Side assaults, extra endeavors excluded in the Metasploit system, and Payload era.

Existing Tools for Penetration Testing

Few of the most well-known apparatuses utilized by security experts for infiltration testing are talked about in this paper. Organize mapper or Nmap is a security scanner device for a PC arrange [5]. This is an open-source programming application essentially used to make a guide of a system and to give a rundown of hosts with related administrations that exist in the system. This apparatus is regularly utilized by experts for performing security reviewing, since the examining of a system may uncover helpless administrations or arrangements. Nmap device can likewise be utilized for system checking and stock. This device is superb versatile

Table 1 Lists some of these tools

Tools/ Techniques	Functions	Availability	Platform	Advantages
Mapper or Nmap [5]	• Security auditing • Network scanning • Port scanning	Freely available as an open source	Linux, Windows, Mac	Excellent scalable
Metasploit [6]	• Work against remote system	Freely available	All versions of Unix and Windows	It is a framework has various functions for security scanning on single platform
Hping [7]	• Remote OS fingerprinting • Security auditing and testing firewalls and networks	Freely available	Windows, Open BSD, Solaris, Mac OS X	Low-level scriptable and idle scanning
SuperScan [8]	• Detect TCP/UDP ports determine which services are running on those ports • Run queries	Freely available	Windows	Possible to access unauthorized open ports

drksujatha23@gmail.com

and this property makes it for examining extensive systems. Another apparatus Metasploit [6] is a system for security testing. This is an abuse structure that gives a few instruments, utilities, and scripts to execute and create misuses against focused remote framework. A variety of different techniques and tools are available for penetration testing. Table 1 lists some of these tools.

3 Testing Work Flow

In this work, we concentrated different automatized devices for infiltration testing. By breaking down the conduct of various instruments, a typical way to deal with automatized entrance testing developed. The strategy took after by these devices comprises three principle stages:

- First, we have to check machines in the system under test to gather all conceivable data.
- Second, we have to recognize vulnerabilities of these hosts by coordinating the aftereffects of the main stage, i.e., check with passages in a powerlessness database.
- In the third stage, it abuses powerlessness to access for a specific asset.

It is hard to discover all vulnerabilities utilizing computerized instruments. There is some powerlessness which can be recognized by manual sweep as it were. Entrance analyzers can perform better assaults on application in view of their abilities and learning of framework being infiltrated. The techniques like social designing should be possible by manual testing as it were. Manual testing process incorporates outline, business rationale with code confirmation.

In the following segment, instruments' methodology will be contrasted and the activities physically performed by an entrance analyzer in a creation domain, with the objective of comprehension the distinctions that make manual testing the favored arrangement in such situations.

4 Conclusion

To consolidate programming testing methods into the advancement, procedure was driven by a longing to adequately enhance the nature of programming while offering help for dealing with the entrance testing. Entrance testing is an exceptionally powerful technique to investigate the shortcoming and quality of system frameworks. By utilizing infiltration test in any association offers advantages, for example, secure organization information, organizations frequently take measures to ensure the accessibility, classification, and trustworthiness of information or to guarantee access for approved people. This paper exhibited a review on the

drksujatha23@gmail.com

examination of a few security devices actualizing infiltration testing and manual testing over system. We attempted to demonstrate a strong technique for the best outcome to secure a system utilizing entrance testing. The objective of this review is to explore the consequences of consolidating manual and computerized approach as semi-automatized intermediary security assessment instrument that mechanizes the security testing of system and in the meantime give control of the testing procedure to the test entertainer. This semi-computerized approach is additionally anticipated that would keep up security assessment instrument with the assistance of a security expert is required to dispose of the issues that can come about by utilizing automatized or manual approach alone.

References

1. Bacudio AG, Yuan X, Chu BTB, Jones M (2011) An overview of penetration testing. Int J Netw Secur Appl (IJNSA) 3(6)
2. Farkhod Alisherov A, Feruza Sattarova Y (2009) Methodology for penetration testing. Int J Grid Distrib Comput 2(2)
3. Shravan K, Neha B, Pawan B (2014) Penetration testing a review. Int J Adv Comput Technol; Compusoft 3(4)
4. Roning J, Laakso M, Takanen A, Kaksonen R (2002) Protos-systematic approach to eliminate software vulnerabilities. https://www.ee.oulu.fi/research/ouspg/
5. Potter B, McGraw G (2004) Software security testing. Secur Privacy IEEE 2(5):81–85
6. Bhattacharyya D, Alisherov F (2009) Penetration testing for hire. Int J Adv Sci Technol 8
7. Smith B, Yurcik W, Doss D (2002) Ethical hacking: the security justification redux. In: Proceedings of 2002 international symposium on technology and society, 2002 (ISTAS'02). IEEE
8. Klevinsky TJ, Laliberte S, Gupta A (2002) Hack IT: security through penetration testing. Addison-Wesley Professional

drksujatha23@gmail.com

Air Pollution Concentration Calculation and Prediction



Jyoti Gautam, Arushi Gupta, Kavya Gupta and Mahima Tiwari

Abstract With the onset of the industrial revolution, the environment is going through severe pollution leading to biological imbalance. The intensity of air pollution in the world has increased at such an alarming rate that it is the need of the hour to determine the changes in the pollution pattern. Air quality dispersion modeling can be done through preferred and recommended models, the most efficient being Eulerian grid-based model. The objective of the paper is to formulate the concentrations of air pollutants using Eulerian model. Various existing methods of prediction work on the basis of models result in satisfactory outcomes but with some certain loopholes. This paper involves methods of predicting pollutant concentration and air quality using machine learning. The data of different sites of Delhi are collected, and the pollutant contributing maximum to the pollution is elucidated using machine learning based methods. Further solutions can be identified to reduce these pollutants.

1 Introduction

Air pollution is growing at a vast rate due to globally increasing industrial development. Industries have come out to be one of the major contributors for increasing the pollution levels and bringing about a drastic change in the pollution pattern. Transportation system even though after being efficient is continuing to cause

J. Gautam (✉) · K. Gupta · M. Tiwari
EasyChair, Noida, India
e-mail: jyotig@jssaten.ac.in

K. Gupta
e-mail: kavya.carmel.15@gmail.com

M. Tiwari
e-mail: triviya2322@gmail.com

A. Gupta · K. Gupta · M. Tiwari
JSS Academy of Technical Education, Noida, Uttar Pradesh, India
e-mail: arushi2410@gmail.com

© Springer Nature Singapore Pte Ltd. 2019

V. S. Rathore et al. (eds.), *Emerging Trends in Expert Applications and Security*,
Advances in Intelligent Systems and Computing 841,
https://doi.org/10.1007/978-981-13-2285-3_30

drksujatha23@gmail.com

pollution, which is accelerating day by day. The proliferating number of vehicles and population is leading the world to a more pollution intense zone, which is causing many types of health hazards as well as decreasing natural resources. The emergent need for a better future is to determine and control the production of pollution from all sources. Air pollution is one of the most dangerous forms of pollution, which needs to be handled immediately as it kills nearly 7 million people annually. Secondary pollutants due to their reactive nature are the leading patrons for all problems and need to be dealt with.

The focus of our work is to develop a mathematical model in such a manner that it calculates the concentration of secondary air pollutants. The base model for our research is the Eulerian grid model. This is a very efficient model also used in the globally accepted THOR model for air quality prediction. The paper outlines the major differences based on certain parameters between the three popular models: Gaussian model, Lagrangian model, and Eulerian model.

Machine learning gives us the ability to learn from experiences rather than being programmed. The paper involves the study of the dataset of the concentrations of air pollutants at certain sites of Delhi, which is used to predict the maximum pollution causing pollutants and to determine air quality using algorithms based on machine learning.

2 Related Work

The most recent work ([1], pp. 8–14) in predicting the level of pollution has been done to compare the low and high levels of PM (2.5) from meteorological data using statistical models and machine learning methods. Another study ([2], pp. 48–54) enlists the impact of mesoscale type wind in a grid-based environment on the pollutants released from a line source which concludes that wind aggravates the concentrations of pollutants for stable conditions as well as neutral conditions.

Different learning techniques can be applied to check the accuracy of prediction of pollutants. A globally accepted model named THOR ([3], pp. 117–122) includes models which can operate on different applications and scales. This model can be used to predict the weather and pollution levels up to 3 days. It helps in emission reduction and pollution management.

Several research works incorporate the differences between various air pollution dispersion models which enable us to determine the most efficient and feasible model. Models such as Gaussian dispersion model ([4], pp. 216–226), Eulerian grid model, and Lagrangian model [5] can be used for prediction of concentration. Few Gaussian dispersion models can be used to deal with secondary pollutants but Lagrangian and Eulerian act as better counterparts because of better statistical accuracy and time consumption. Eulerian stands out to be the better among the three due to its high accuracy (Table 1).

drksujatha23@gmail.com

Table 1 Comparison of various models

Parameter	Gaussian	Lagrangian	Eulerian
Type	Static, dynamic	Dynamic	Static
Artificial diffusion (AD)	Possible	Not possible	Possible
Computation cost	Lowest	Low	High
Accuracy	Not accurate	Not accurate	Accurate as it uses high grid resolution
Dense areas	Low accuracy	Low accuracy	High accuracy
Statistical accuracy	Average	Not good	Very good
Time consumption	Less	More	Less
Reliability	Unreliable	Reliable	Reliable

$$\frac{\partial C}{\partial t} = -\mathrm{U} \cdot \nabla C + \mathrm{D}\nabla^2 C + \mathrm{S} \tag{2.1}$$

where C is the concentration of pollutant in the air; U is the wind speed vector of the components u, v, and w; D is the diffusion coefficient; S is the source and sink of pollutant; ∇ is the Gradient operator; and ∇^2 is the Laplacian operator.

With the application of certain assumptions (Goyal et al. 2011, pp. 105–114), the above equation is transformed to

$$U\frac{\partial C}{\partial x} = \frac{\partial}{\partial y}\left(K_y \frac{\partial C}{\partial y}\right) + \frac{\partial}{\partial z}\left(K_z \frac{\partial C}{\partial z}\right) \tag{2.2}$$

where x, y, and z are the coordinates along the wind, the crosswind, and the vertical directions, respectively. C is the mean concentration of the pollutants, and U is the mean speed of the wind for downward direction. Ky and Kz are eddy diffusivities in crosswind and vertical directions for the pollutants, respectively.

3 Methodology

A dataset is prepared consisting of the concentration of pollutants in Azure machine learning understandable format. This data is about the average concentration of various secondary pollutants from different areas of Delhi as well as the date and time of its release. Comma-Separated Value (CSV) is a tabular form of data representation that saves plaintext data separated by commas. Azure comprehends with this type of format enabling easy presentation of data.

drksujatha23@gmail.com

Data preprocessing techniques to make the dataset consistent need to be applied in order to make raw, noisy, and unreliable data clean, and normalize the results thus obtained. Z-score algorithm is a technique for normalization that transforms the value of all variables to common scale of zero and standard deviation of the variables by avoiding aggregation distortions that are found due to different mean of variables. Calculation of z-score is performed by the following equation:

$$Normalized\,(xi) = \frac{x(i) - \bar{x}}{s(x)} \tag{3.1}$$

where Normalized (xi) is the normalized value for the ith record of variable x, and x (i) is the value of x for record i,

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} xi \tag{3.2}$$

And

$$s(x) = \sqrt{\frac{\sum_{i=1}^{n}(xi - \bar{x})^2}{n}} \tag{3.3}$$

n is the total number of the values of the variable.

Permutation feature importance is a technique that enables us to simulate the effect of a set of features of the data by computing performance scores for the model based on shuffling of values of the variables randomly. These scores represent the change in the accuracy of the trained model if the value of that variable changes. This helps to determine the pollutant which is contributing most to the pollution. The main ideology behind this algorithm is the one similar to the feature randomization process implemented in the random forests given by Brieman. IT follows the main assumption that the permuted values of important features result in a significant change in the trained model as compared to the less important features.

The model is trained using the decision forest regression algorithm. Decision trees are created which work as nonparametric models and perform a sequence of tests for every instance of the variable until the complete tree is traversed and a goal or decision is found. These are computationally efficient, also have limited memory usage and are also resilient for noisy data. A decision forest is created using large number of decision trees, followed by resampling of the data using bagging and a large number of splits are specified per node. Several decision forest algorithms are checked for the maximum accuracy of the result, thus giving us the major pollutant contributor of pollution.

Further to evaluate the performance, certain measures (Li et al. 2016, pp. 22408–22417) have been used to determine the accuracy of the prediction:

drksujatha23@gmail.com

$$RMSE = \sqrt{\frac{\sum_1^n (Oi - Pi)^2}{n}} \tag{3.4}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |Oi - Pi| \tag{3.5}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n} \frac{|Oi - Pi|}{Oi} \tag{3.6}$$

where RMSE stands for Root-Mean-Square Error, MAE stands for Mean Absolute Error, and MAPE stands for Mean Absolute Percentage Error. Also, Oi is the set of observed values, Pi is the set of predicted values, and n is the total number of values.

4 Conclusion

In this project, a mathematical model based on Eulerian grid and machine learning is developed for pollutant concentration calculation and air quality prediction. The project outlines the major differences based on certain parameters between the three popular models: Gaussian model, Lagrangian model, and Eulerian model. Regression-based algorithms of machine learning are used to achieve maximum accuracy in prediction of the maximum pollutant contributor of air pollution in areas of Delhi. Concentrations of pollutants such as CO, O_3, and $PM_{2.5}$ have been taken for two years 2016–18 (source: CPCB or Central Pollution Control Board),

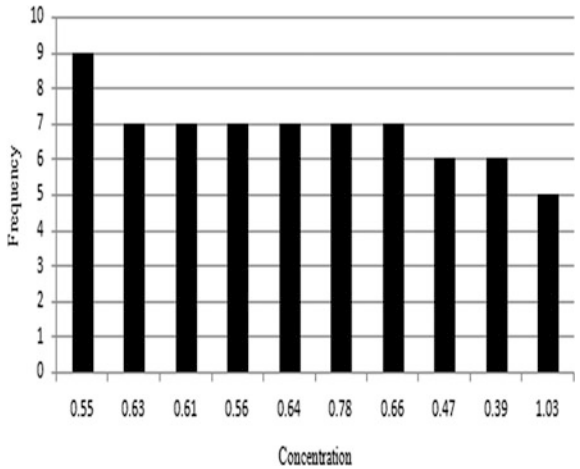


Fig. 1 Concentration of CO

drksujatha23@gmail.com

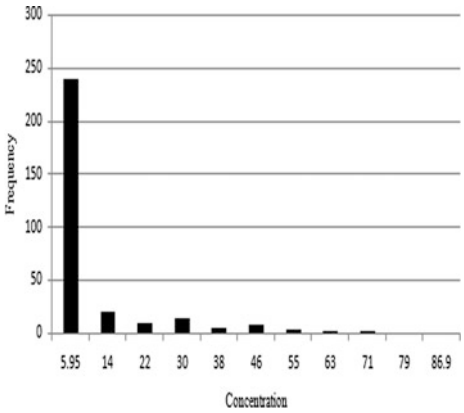


Fig. 2 Concentration of O_3

and the frequency of their concentrations over this time span is represented in Figs. 1, 2, and 3, respectively. These figures represent the variations in the concentrations of these pollutants over a certain span of time by depicting their respective frequencies for a particular concentration.

Further, certain parameters such as Relative Humidity (RH), Temperature (Temp), Primary pollutant (SO_2), and other secondary pollutants have been analyzed as to how they affect or bring about any changes in the pollution caused by $PM_{2.5}$. This analysis is done using permutation feature importance algorithm in Azure ML Studio of Microsoft (Fig. 4). It can be clearly depicted that SO_2 has less or no effect on the concentrations of $PM_{2.5}$, whereas all the other parameters affect the pollution done by $PM_{2.5}$ (Table 2).

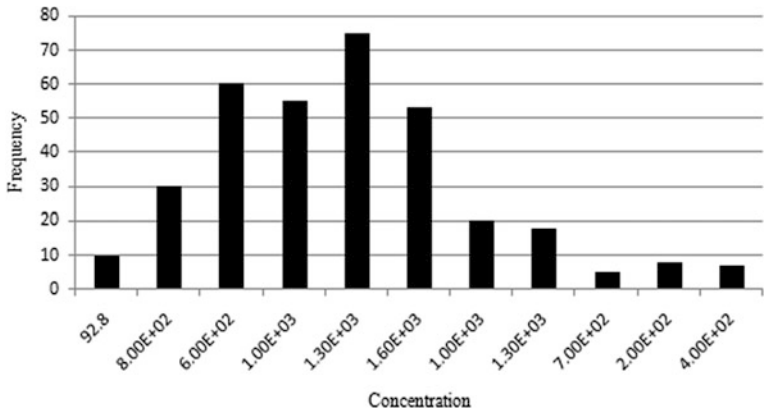


Fig. 3 Concentration of $PM_{2.5}$

drksujatha23@gmail.com

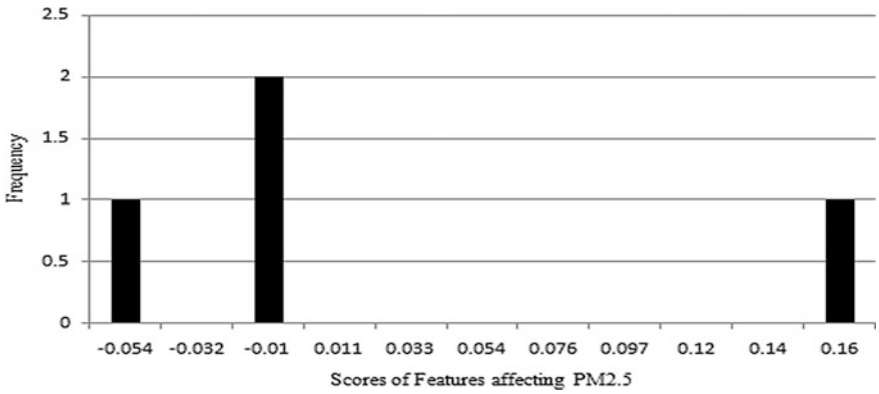


Fig. 4 Score of features affecting concentration of $PM_{2.5}$

Table 2 Feature importance according to their score

Feature	Score
RH	0.162057
SO_2	0
Temp	−0.00008
O_3	−0.053551

Based on the research, further work can be performed for predicting the concentration level of various pollutants such as NO_2, NO, CO_2, etc. The findings listed above help us to prognosticate the future values as well as determine the error between the predicted value and observed value. The root-mean-square error, the mean absolute error, and the mean absolute percentage error can be used as the parameters for the comparison stated above.

References

1. Kleine Deters J, Zalakeviciute R, Gonzalez M, Rybarczyk Y (2017) Modeling $PM_{2.5}$ urban pollution using machine learning and selected meteorological parameters. Accepted 11 May 2017
2. Krishna S, Lakshminarayanachari K, Pandurangappa C (2017) Mathematical modelling of air pollutants emitted from a line source with chemical reaction and mesoscale wind. Int J Sci Eng Res 8(5)
3. Brandt J, Christensen JH, Frohn LM, Zlatev Z (2002) Operational air pollution forecast modelling using the THOR system. Department of Atmospheric Environment, National Environmental Research Institute
4. Juodis L, Filistovic V, Maceika E, Remeikis V (2016) Analytical dispersion model for the chain of primary and secondary air pollutants released from point source. Atmos Environ 128:216–226
5. Pillai D, Gerbig C, Kretschmer R, Beck V, Karstens U, Neininger B, Heimann M (2012) Comparing Lagrangian and Eulerian models for CO_2 transport–a step towards Bayesian inverse modeling using WRF/STILT-VPRM

drksujatha23@gmail.com

The NLP Techniques for Automatic Multi-article News Summarization Based on Abstract Meaning Representation



Deepa Nagalavi and M. Hanumanthappa

Abstract The analysis of natural language texts is one of the most important knowledge discovery tasks for any organization. Automated text summarization systems can reduce the size of the text while keeping the important part of the text and desired information. The applications of summarization are summaries of email thread, action items from a meeting, simplifying text by compressing sentences and abstracts of any document, an article, etc. It comes in two ways, single-document summarization and multiple-document summarizations. In single-document summarization technique, given a single document produces abstract, outline. Whereas, with multiple-document summarization technique, given a group of documents produce a list of the content such as a series of news stories on the same event or a set of web pages about some topic or questions. Consequently, there are two ways of doing summarization, an extractive summarization creates the summary from phrases or sentences in the source document and an abstractive summarization express the ideas in the source documents using different words. However, the abstractive summarization methods are very comprehensive to get the abstract meaning of multiple articles and generate the summary. Thus, the text contents are analyzed and extract named entities using Stanford NER tool in different aspects to get abstract meaning of multiple articles. In this study, an abstract summary of article using named entities are presented.

Keywords Natural language processing · E-Newspaper · Summarization Stanford NER

D. Nagalavi (✉) · M. Hanumanthappa
Department of Computer Science and Applications,
Bangalore University, Bangalore, India
e-mail: deepatnagalavi@bub.ernet.in

M. Hanumanthappa
e-mail: hanu6572@bub.ernet.in

© Springer Nature Singapore Pte Ltd. 2019

V. S. Rathore et al. (eds.), *Emerging Trends in Expert Applications and Security*,
Advances in Intelligent Systems and Computing 841,
https://doi.org/10.1007/978-981-13-2285-3_31

drksujatha23@gmail.com

1 Introduction

Text summarization is a booming technique which does the task of creating coherent summaries that state the main purpose of the given textual document. Natural language processing system provides two types of summarization methods one is generic and the other one is query focused summarization. In Generic summarization the system provides the summarized content of document. Whereas, in query-focused summarization it summarizes a document with respect to an information need expressed in user query. It can also be used to create answers to complex questions by summarizing multiple documents instead of giving a snippet for each document, it creates a cohesive answer that combines information from each document.

News article summarization technique generates summaries of multiple articles on the same topic. The natural language processing (NLP) techniques are used for text summarization process. Text summarizers are divided into two categories, linguistic and statistical. The process of analyzing an abstract meaning of sentences and summarize the text by changing sentences are Linguistic summarizers, it is also known as abstractive summarizer. It makes use of the knowledge about the language to summarize articles. However, statistical summarizer also called extractive summarizer selects existing important words, phrases, or sentences using statistical methods. Thus, an extractive summarization method gives an idea about the content of the input article. Additionally, it should satisfy the need of efficient summarization method namely, textuality, significance, and compression constraints [1]. The summarization algorithms are evaluated based on supervised or unsupervised techniques. However, this statistical method is intent to identify the most important areas in the context of words, sentences, and paragraphs, among others, in the input sourced from one or more article documents. Summaries derived from the extraction procedure hold some concatenated sentences expressed precisely as they occur in the articles targeted for summarization. In the extractive summarization procedure, a ruling is arrived at on whether or not a specific sentence ought to be extracted for inclusion in the summary. Search engines, for instance, employ extractive summary generation to realize summaries from web pages.

2 Related Work

A variety of summarization approaches has been proposed in the literature. Multi-article summarization is aimed summarizing information from multiple texts in compact and concise manner. These approaches are designed and developed using the models, such as feature-based model, graph-based ranking model, integer linear programming model and others.

Most existing researches emphasis on extractive summarization methods, in which the important and more frequent sentences are selected from the detailed

drksujatha23@gmail.com

news article set. Therefore, in the literature Zhang et al. [2] proposed a method based on symbolic characteristics and structural information. It is a statistical method to measure the similarity among sentences based on semantic similarity is used in [3–5], while identifying the important terms or sentences. The semantic similarity methods are classified into four categories, an edge counting methods, information content methods, feature-based methods, and hybrid methods.

Anjali and Lobo [6] suggested this novel approach to multi-document summarization that warrants excellent coverage and avoidance of generation of redundant sentences. The input to the query is the group of texts and query. These authors have retained a map list, where every expression with its frequency from the text group is stowed in a map. The technique of query modification is applied as follows: query splitting into tokens and finding the synonym for every token, and if the synonym or tokens in a text collection then the highest frequent add the synonym of the query with highest frequent to query. The terms occurring most frequently among corpus are chosen and added to the query for strengthening of the query. The elements are employed in the computation of sentence score which include title feature, numerical data, cue phrase, sentence length, sentence centrality, sentence position, upper case word, term frequency, sentence similarity, and inverse document frequency. Clustering of the text is done via use of cosine similarity as a means of generating the necessary documents clusters. Thereafter, from each text cluster, clustering of sentences is based on the values of resemblance. Each group's score is then calculated. Sorting of the sentence clusters is then done in the reverse order of the group score. Lastly, for every cluster, the best sentence score is selected and added to the final summary.

The extractive multi-document summarizer was proposed by Amit and Aarati [7]. This method is a graph-based multi-document summarizer that follows the following steps. A set of related texts forms the input. In the first phase, pre-processing of documents is done. The undirected acyclic graph is generated for every text with sentences representing nodes and similarities representing edges. Then, the weighted ranking algorithm is executed to allow generation of significant score for every sentence in the text. Sentences ranking is done based on their respective significant scores. The highest ranking sentences are selected from the summary of every text. In the second phase, all the single summary of every text is assembled to form a single text. In the last phase, the above process is employed in combining document, thus forming the last extractive summary.

3 Multi-article Summarization

Single article summary gives the brief information of particular article from one newspaper. However, the summary of multiple articles from multiple newspapers which are discussed on same topic needs to be summarized to optimize the time reading multiple newspapers for one event. Thus, multi-document summarization approaches are determined evaluated with the Document Understanding

drksujatha23@gmail.com

Conferences, [1] conducted annually by NIST. For multi-document summarization, the highly skilled analysis criteria are considered. It also considers other evaluation criteria bring out the summary from more general to more specific aspects with good readability. It needs more intelligence to write summary in different words and patterns. Comparatively, it is quite difficult to generate similar summaries because it needs to incorporate artificial intelligence, semantic representation, natural language generation and designing a conclusion on the basis of evidence and reasoning.

Article summarization shortens the source article texts, while considering the key aspects and various other factors to organize a sentence. Therefore, the proposed model combines the features of both extractive and abstractive approach. It generates the summary in two or three line based on named entities. Thus, the work is divided into three phases, first is named entity recognition model is used to identify and extract the entities, then generate the summarized sentence using entities as keyword and finally, combines the summary of multiple articles of same topic. However, the articles are first preprocessed which consists of the NLP preprocessing methods such as cleaning up the text and tokenizing the article content by locating word boundaries. Later, analyze the sentences and identify the dependency using the Stanford Parser. Mainly, it recognizes Named Entities using Stanford NER. Thus, the stanford parser identifies the terms and the dependency of the sentences while tagging the whole document with named entities called Stanford NER tags. It also extracts the headline and newspaper name from the document. The work flow of summarizer is derived in Fig. 1. According to this figure when user enters a query to search for an article from multiple newspapers the query processor searches articles and sends to the summarizer. The summarizer analyzes the article text and identifies the named entities, the relationships and typed dependency to extract the important information. Later sentences are

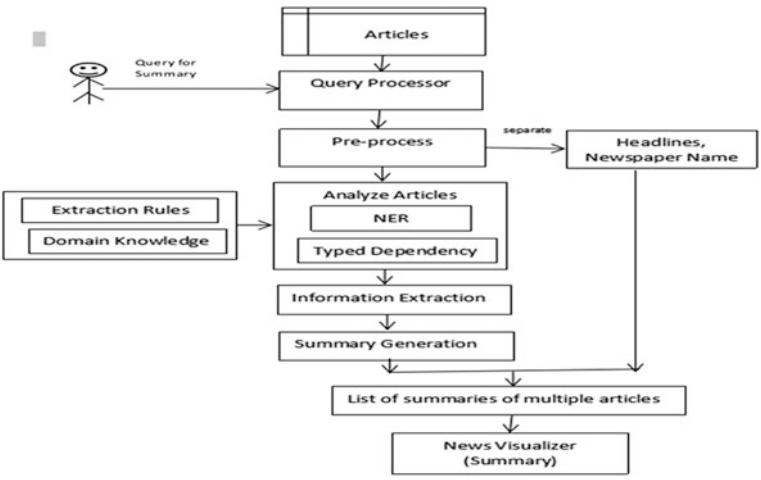


Fig. 1 Work flow of article summarizer

drksujatha23@gmail.com

generated based on the entities extracted. Finally, the summary of multiple articles are present in news visualizer.

Stanford NER is a Java implementation of a Named Entity Recognizer. It identifies the entities and labels the sequences of words in a text such as the names of things, person, and company names [8]. Later, extract the labeled entities using feature extractors with multiple options. Admittedly, compare to other NER tool, the Stanford NER is a very good named entity recognizers for English. It gives better result particularly for the seven classes. It also comes with multiple options as such the NormalizedNamedEntityTagAnnotation normalized value stored in the sequence. The rule-based NER, the RegexNER annotator [8, 2] are also used to identify the entities and the dependency value. The named entity recognizer by default recognizes named (Person, Location, Organization, Misc), numerical (Money, Number, Ordinal, Percent), And Temporal (Date, Time, Duration, Set) entities (12 classes). Adding the regexner annotator and using the supplied RegexNER pattern files adds support for the fine-grained and additional entity classes Email, Url, City, State_Or_Province, Country, Nationality, Religion, (job) Title, Ideology, Criminal_Charge, Cause_Of_Death (11 classes) for a total of 23 classes. The tool is recognizing the entities using a combination of three CRF sequence taggers and rule based system. The tool uses various corpora, including CoNLL, ACE, MUC, and ERE corpora.

The news articles contain information answering all six WH's questions based on the effect of the event. Thus, such information are extracted by analyzing the text from news article. The analyzer works well in five steps. In the first step, the named entity recognition model is used to identify the key entities, which gives the information of location. Here location means it tries to identify where the event happened or where it is filed. In the second step, the dependency or the relationship with previously identified information is identified in the sentences, the information consists of the main event of the article. In the next step, by using the typed dependency method the cause of the event is identified. The regexner annotator is also used to analyze the sentence and extracting the information. Therefore, the dependency relations and key entities are generated by using the Stanford parser.

The information extracted from the previous step is used to generate the brief summary. The extracted keywords are framed and generate the sentence. It follows the straightforward sentence generation design patterns [9]. The Information extraction module selects important keywords that serve as noun phrase head, and its number, modifiers, and specifier for sentence generation. Multi-article summary is presented in table form which presents the summary of each article. The summary consists of the theme of the event. Exact similar summaries are eliminated from the table however dissimilar summaries are presented in table with the newspaper name. The designed search engine in the proposed system search for the articles based on query and summarizes each article and combines the result and present to the user. The table consists of form which newspaper, the headline, the summary and the link to read the original article.

drksujatha23@gmail.com

4 Summary Evaluation

The proposed approach is evaluated by using the dataset provided by DUC 2002 document. Generally, DUC provides a standard corpus which is normally used in text summarization task. Especially, it provides ROGUE metric to evaluate the summary, which provides multiple properties that make it unsuitable for evaluating abstractive summaries. One of the reasons why ROGUE like metrics might never become suitable for evaluating abstractive summaries is its incapability of knowing if the sentences have been restructured. A good evaluation metric should be one where we compare the meaning of the sentence and not the exact words.

The proposed approach is evaluated by employing both qualitative and quantitative evaluation metrics. The news articles from different sources are taken into consideration for the experimental analysis. Multiple articles are summarized based on the information extracted using NER tool. The tool identifies most of important entities which are enough to convey the summarized information to the user. According to DUC guidelines the summary should provide all the required information in brief [2, 10]. Thus, the summary is evaluated whether it is answering the requirement of a good summary or not. Consequently, a set of news articles are taken into consideration to evaluate the generated summary which is rated by humans. The summaries are evaluated by humans and rate the results based on the parameters information content, reader satisfaction, summary length, grammatical correctness. Table 1 shows the experimental result. Table shows the ratings given by 10 persons out of 5. It gives the summary of evaluation matrix of proposed work amongst the 50 news articles, it shows the average ratings of the summaries generated by extractive and abstractive methods given by different persons. The summary is evaluated in terms of the criteria given in DUC 2002. Also the proposed approach is compared with the extractive method and observed that the proposed abstractive approach gives better result. However, the length of the summary compare to the original article observed to be less than one-third.

Table 1 Experimental results

Persons	Human written	Extractive approach	Proposed approach
1	4.0	2.0	3.5
2	3.5	1.2	3.3
3	3.1	1.6	3.5
4	3.4	2.0	3.0
5	4.2	1.9	4.0
6	2.9	1.4	3.0
7	4.0	1.5	3.2
8	3.8	1.2	2.5
9	3.0	2.0	3.4
10	3.5	2.5	3.2

drksujatha23@gmail.com

5 Conclusion

In this paper, the automatic multi-article news summarization approach based on abstract meaning representation is implemented. The resultant summaries are informative to give brief information about an event and competitively short in size. The proposed model is implemented based on multiple parameters such as named entity, its relationship with sentence, typed dependency, and many more to generate a structured summary. Thus, the Stanford NLP parser tools are used to extract the essential information from the text. However, abstractive summarization method compare to extractive method anticipate very challenging task, this summarizer works on a wide variety of domains varying between international news, politics, entertainment, and so on. Another useful feature is that it summarizes multiple article of same topic from different newspaper and present in the list. The proposed approach is as good as the technologies similar to artificial intelligence. As it is analyzing the texts, pinpointing key concepts, and produces instant summaries.

References

1. Rush AM et al (2015) A neural attention model for abstractive sentence summarization. arXiv:1509.00685v2 [cs.CL], 3 Sept 2015
2. Zhang J, Sun Y, Wang H, He Y (2011) Calculating statistical similarity between sentences. J Converg Inf Technol 6(2):22–34
3. Sahoo D et al (2016) Aspect based multi-document summarization. In: International conference on computing, communication and automation (ICCCA2016). ISBN:978-1-5090-1666-2/16 ©2016 IEEE
4. Ramanujam N et al (2016) An automatic multidocument text summarization approach based on Naive Bayesian classifier using timestamp strategy. e Sci World J 2016, Article ID 1784827, 10 p. Hindawi Publishing Corporation. http://dx.doi.org/10.1155/2016/1784827
5. Nallapati R et al (2007) SummaRuNNer: a recurrent neural network based sequence model for extractive summarization of documents. In: Proceedings of the 31st AAAI conference on artificial intelligence 2017
6. Deshpande AR et al (2013) Text summarization using clustering technique. Int J Eng Trends Technol (IJETT) 4(8)
7. Zore AS et al (2014) Extractive multi document summarizer alogorithim. (IJCSIT) Int J Comput Sci Inf Techno 5(4):5245–5248
8. Jhalani R et al (2017) An abstractive approach for text summarization. Int J Adv Comput Eng Netw 5(1). ISSN:2320-2106
9. Shimpikar S et al (2017) A survey of text summarization techniques for Indian regional languages. Int J Comput Appl (0975–8887) 165(11)
10. Belkebir R, Guessoum A (2016) Concept generalization and fusion for abstractive sentence generation. Expert Syst Appl 53(2016):43–56. Elsevier Ltd
11. Bagalkotkar A et al (2013) A novel technique for efficient text document summarization as a service. In: 2013 third international conference on advances in computing and communications, 978-0-7695-5033-6/13, IEEE. https://doi.org/10.1109/icacc.2013.17
12. Liu F et al (2015) Toward abstractive summarization using semantic representations. Carnegie Mellon University Research Showcase @ CMU, Computational Linguistics: Human Language Technologies, pp 1077–1086

drksujatha23@gmail.com

13. Garje GV et al (2016) Generating multi-document summarization using data merging technique. Int J Comput Appl (0975–8887) 138(6)
14. Damonte M et al (2017) An incremental parser for abstract meaning representation. In: From the proceedings of EACL 2017, Valencia, Spain
15. Gupta V et al (2012) An statistical tool for multi-document summarization. Int J Sci Res Publ 2(5). ISSN: 2250-3153
16. Munot N et al (2014) Comparative study of text summarization methods. Int J Comput Appl (0975–8887) 102(12)
17. Shinde RD et al (2014) Enforcing text summarization using fuzzy logic. Int J Comput Sci Inf Technol 5(6):8276–8279
18. Yang S et al (2017) KeyphraseDS: automatic generation of survey by exploiting keyphrase information. Neurocomputing 224(2017):58–70. 0925-2312/ © 2016 Elsevier B.V
19. Takase S et al (2016) Neural headline generation on abstract meaning representation. In: Proceedings of the 2016 conference on empirical methods in natural language processing, pp 1054–1059, Austin, Texas, November 1–5. © 2016 Association for Computational Linguistics
20. Ma S et al (2016) An unsupervised multi-document summarization framework based on neural document model. In: Proceedings of COLING 2016, the 26th international conference on computational linguistics: technical papers, pp 1514–1523, Osaka, Japan, December 11–17 2016
21. Sunitha C et al (2016) A study on abstractive summarization techniques in Indian languages. Procedia Comput Sci 87:25–31. 1877-0509 © 2016 by Elsevier B.V
22. AL-Khassawneh YA et al (2016) Sentence similarity techniques for automatic text summarization. J Soft Comput Decis Support Syst 3(3):35–41
23. Cao Z et al (2016) TGSum: build tweet guided multi-document summarization dataset. In: Proceedings of the thirtieth AAAI conference on artificial intelligence (AAAI-16) Copyright © 2016, Association for the Advancement of Artificial Intelligence

drksujatha23@gmail.com

An Analysis of Load Management System by Using Unified Power Quality Conditioner for Distribution Network



D. Jayalakshmi, S. Sankar and M. Venkateshkumar

Abstract This paper focused to designing and control of unified power quality conditioner (UPQC) for improving the load enhancement. In this paper, the authors presented the modeling of UPQC-based power system network for improving the voltage profile and load enhancement in a radial distribution power network under various power system faults. The objective of this paper is to improve the voltage profile and load enhancement using an intelligent controller based UPQC device. The operation of UPQC devices is to be analysis using intelligent controller at various fault conditions. The intelligent controller compares the power system parameters such as voltage and phase angle with the reference value and it will generate the triggering pulses for a voltage source converter of UPQC system. The proposed model will be simulated in MATLAB environment. The simulation results are evaluated with IEEE standards and compare to existing models for strong impact of the proposed model.

Keywords UPQC · Fuzzy · Voltage improvement · Load enhancement and Matlab

1 Introduction

Bound together power quality conditioners (UPQCs) comprise of consolidated arrangement and shunt dynamic power channels (APFs) for concurrent pay voltage and current unsettling influences and responsive power. They are relevant to control

D. Jayalakshmi · S. Sankar · M. Venkateshkumar
St Peter's University, Chennai, India
e-mail: Jayaeec28@gmail.com

D. Jayalakshmi · S. Sankar · M. Venkateshkumar
KKC Institute of Technology and Engineering, Puttur, India

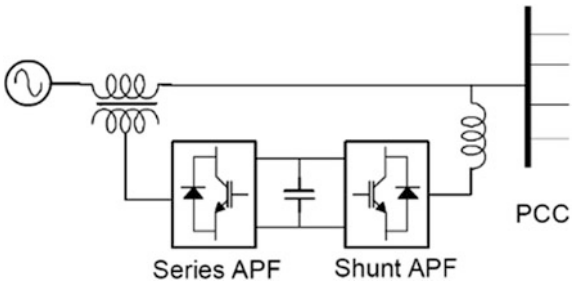
D. Jayalakshmi · S. Sankar · M. Venkateshkumar (✉)
Department of EEE, Aarupadai Veedu Institute of Technology, Chennai, India
e-mail: venkatmme@ieee.org

© Springer Nature Singapore Pte Ltd. 2019

V. S. Rathore et al. (eds.), *Emerging Trends in Expert Applications and Security*,
Advances in Intelligent Systems and Computing 841,
https://doi.org/10.1007/978-981-13-2285-3_32

drksujatha23@gmail.com

Fig. 1 UPQC—circuit model



conveyance framework and is associated at the point of common coupling (PCC) of burdens that create consonant streams. Assorted topologies have been proposed in the writing for UPQCs in single-stage setups, i.e., two IGBT half extensions [1] or multilevel topologies [2], however this paper concentrates on the usually utilized general structure delineated in Fig. 1 [3]. As can be seen, the power converters share a dc-bus and, contingent upon their functionalities, utilize a separation transformer (arrangement APF) or an inductance (shunt APF) as voltage or current connections.

The arrangement APF must repay the source voltage unsettling influences, for example, sounds, plunges or over-voltages, which may fall apart the operation of the neighborhood stack while the shunt APF weakens the bothersome load current segments (consonant streams and the central recurrence part which adds to the receptive load control). Also, the shunt APF must control the dc-bus voltage to guarantee the compensation capacity of the UPQC [4]. These functionalities can be completed by applying various control procedures which can work in the time area, in the recurrence space or both [5]. Time space techniques, for example, pq- or dq-based strategies [6–8], permit the quick compensation of time-variation unsettling influences however make more intricate their particular pay. In this sense, recurrence area strategies are more adaptable however their dynamical reaction is slower.

2 Functional Structure of UPQC

Practical structure of UPQC the essential functionalities of an UPQC controller are represented in Fig. 2. The voltage remuneration (vC∗) and current infusion (iC∗) reference signals, required for compensation intentions, are assessed from the prompt estimations of the source voltage (vS), the dc-transport voltage (vdc) and the heap current (iL). These reference signals are contrasted with the deliberate criticism signals v1 and i2 and connected to the decoupled voltage and current controllers, which guarantee that the pay signals compare to the reference ones. The door signs of the power converters are acquired by applying beat width modulators to the controller yields.

drksujatha23@gmail.com

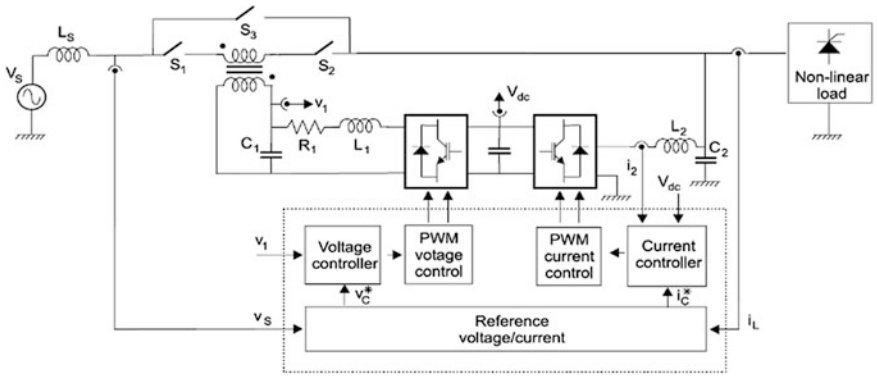


Fig. 2 Functional structure of UPQC the basic functionalities of a UPQC controller

The power converters switch at high recurrence producing a PWM yield voltage waveform which must be low-pass separated (L1, R1, and C1 if there should arise an occurrence of arrangement APF and L2 y C2 for the shunt APF). Switches S1, S2, and S3 control the pay status of the UPQC. The voltage controller can be actualized in three ways. Criticism structures permit a decent stationary reaction while forward structures produce fast reactions amid voltage homeless people. Encourage forward structures permits the two practices being more utilized [9]. The age of the reference flag depends emphatically on the remuneration targets: voltage dips, over-voltages or voltage harmonics. The rms estimation of the network voltage can be measured to distinguish voltage dips and over-voltages, once identified, the PLL used to synchronize the compensation standard must be solidified (not connected to the voltage motion) to keep up the past stage. At the point when the heap voltage harmonic is the compensation objective, a tedious controller can be connected to alleviate the impact of all voltage harmonic [9]. For this situation the reference signal is created inside the voltage controller and does not permit specific consonant pay, both in symphonious request and consonant greatness. Diverse methodologies have been proposed for current control of network associated voltage source converters. Hysteresis controllers are actualized by methods for straightforward simple circuits be that as it may, as disadvantage, the range of the yield current is not confined, which confounds the yield channel plan [10]. PI controllers have been broadly connected at the same time, because of their limited pick up at the central network recurrence, they can present enduring state blunders. This can be understood by methods for summed up integrators [11]. Fluffy rationale and counterfeit fuzzy has been additionally proposed as present controllers in the event of various symphonious frequencies in the reference current signals [12–15].

drksujatha23@gmail.com

3 Design of Fuzzy Controller for UPQC

The proposed fuzzy logic controller has been designed for better controller of UPQC and improves the power quality as shown in Fig. 3. The fuzzy controller has two inputs such as Id and Iq measured Current. The input membership functions are designed based on trapezoidal methods are presented in Figs. 4 and 5. The fuzzy controller has two outputs such as Id and Iq regulated Current. The output membership functions are designed based on trapezoidal methods are presented in Fig. 6a, b. The defuzzification of proposed controller has been designed using

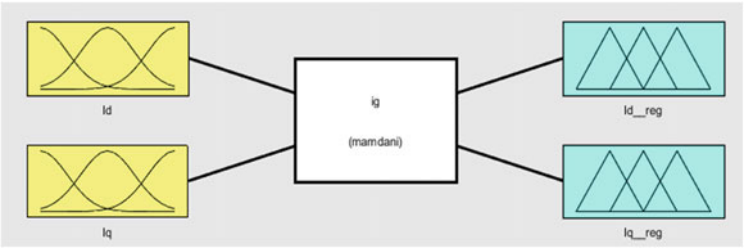


Fig. 3 Fuzzy controller structure for UPQC

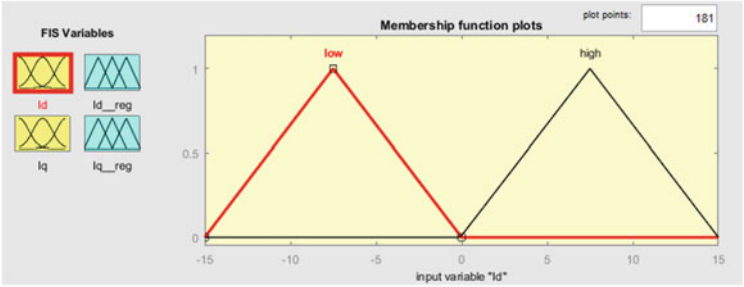


Fig. 4 Input fuzzy membership functions for UPQC (Id measured Current)

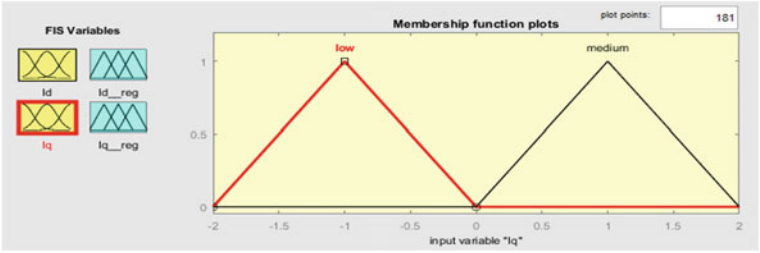


Fig. 5 Input fuzzy membership functions for UPQC (Iq measured Current)

drksujatha23@gmail.com

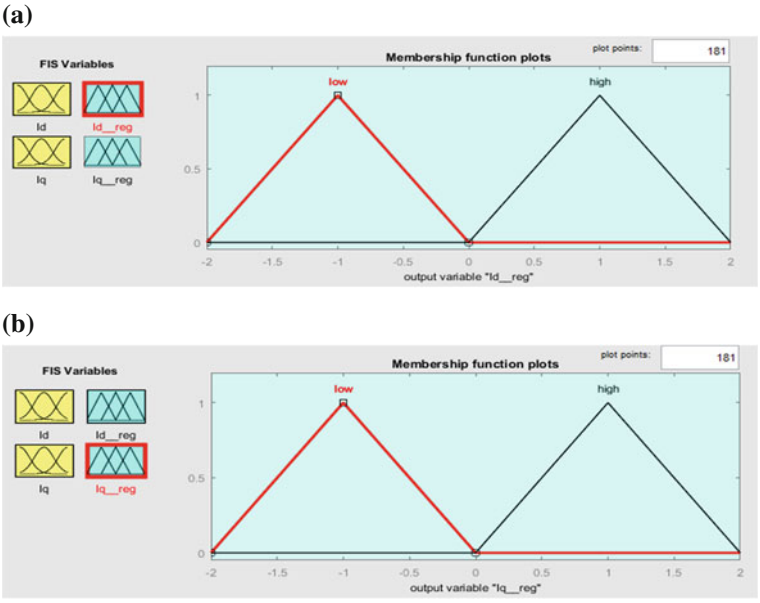


Fig. 6 **a** Output fuzzy membership functions for UPQC (Id regulated Current) **b** Output fuzzy membership functions for UPQC (Iq regulated Current)

center of gravity. Finally, the fuzzy interface rules are formed based on input membership function and presented in Fig. 7.

The simulation model of UPQC with fuzzy controller and power system model are presented in Fig. 8a, b. Figure 9 is represented the fuzzy controller in UPQC system. The fuzzy controller has generated duty cycle of the PWM pulse to UPQC based on grid voltage and the reference voltage during fault conditions. The proposed two parallel operating power system models are simulated and apply the

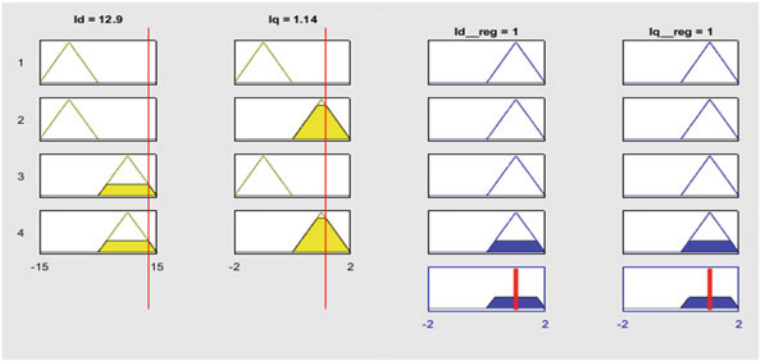


Fig. 7 Fuzzy interference rules for UPQC controller

drksujatha23@gmail.com

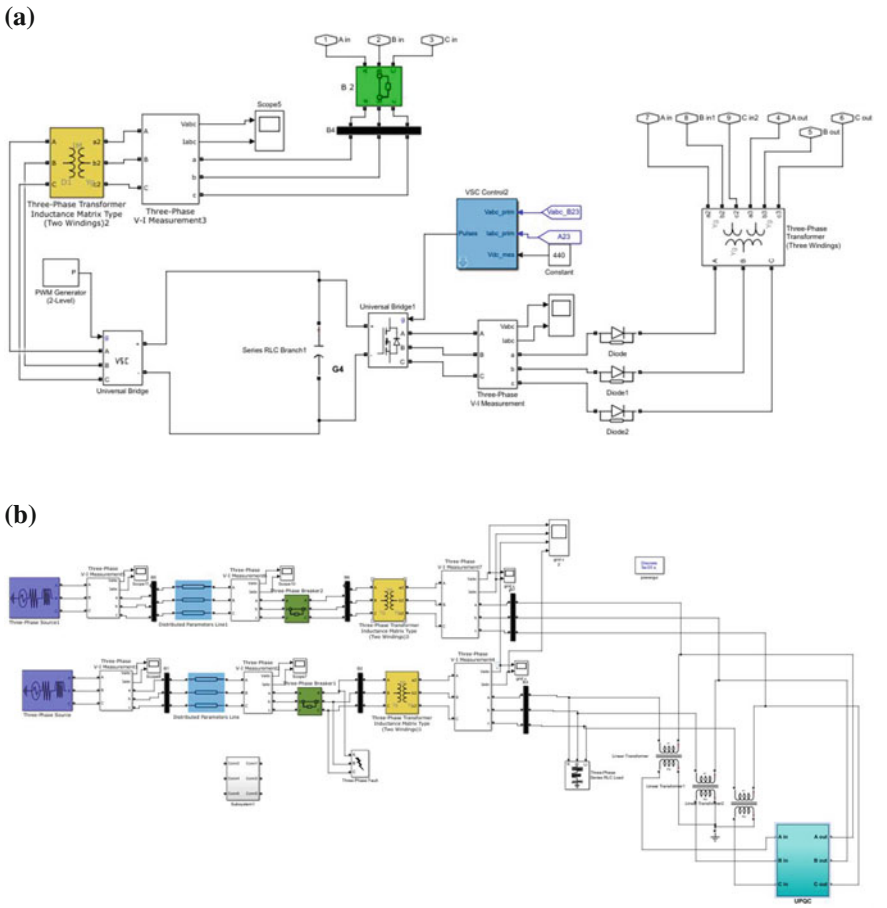


Fig. 8 **a** Simulation model Of UPQC with fuzzy controller. **b** Simulation model Of UPQC with fuzzy controller in distributed power network

different fault conditions in power system 2. The proposed fuzzy logic controller based UPQC will improve the power quality and controller fault in-between power system [16]. The three-phase fault and its THD values 0.93% are shown in Fig. 10 and Fig. 11 respectively. The Line to ground faults, it is introduced in phase A and Ground as shown in Fig. 12 and its THD value is 0.65% presented in Fig. 13. The line to ground faults occurs at 0.2–0.4 s and after 0.4 s the system settled under normal condition. The Line to Line faults it is introduced in phase A and Phase B as shown in Fig. 14 and its THD value is 0.74% presented in Fig. 15. The line to line faults occurs at 0.2–0.4 s and after 0.4 s the system settled under normal condition. The Double Line to ground faults, it's introduced in phase A, Phase B and Ground as shown in Fig. 16 and its THD value is 0.87% presented in Fig. 17. The double line to ground faults occurs at 0.2–0.4 s and after 0.4 s the system settled under

drksujatha23@gmail.com

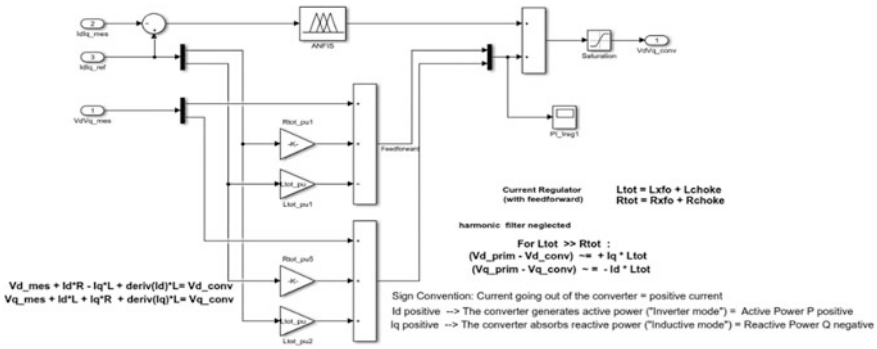


Fig. 9 Fuzzy logic controller for UPQC

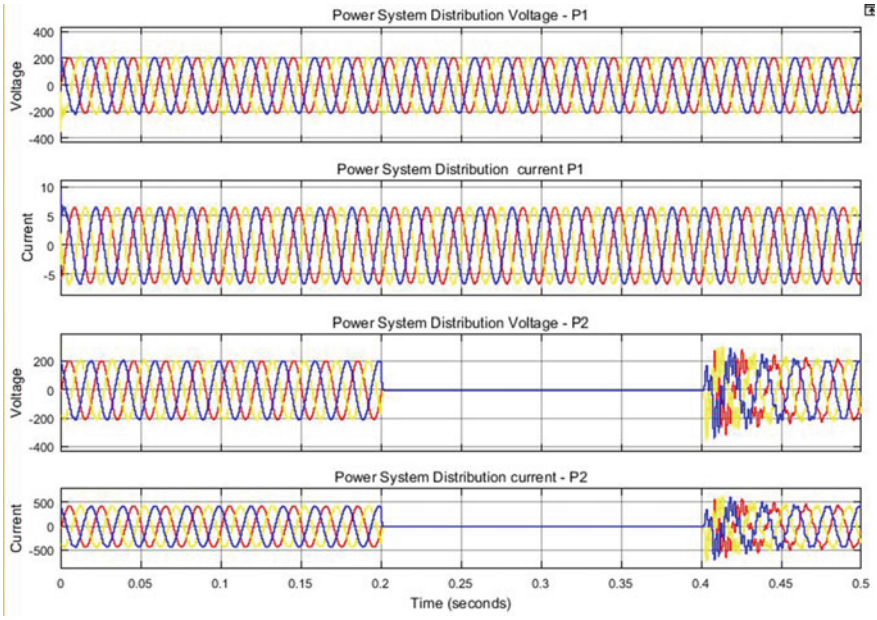


Fig. 10 Three-phase fault with UPQC + fuzzy controller

normal condition. The above analysis is clearly represented after a fault occurs the system did not lose the stability using a fuzzy logic controller based UPQC device. The proposed system has effectively maintained the power system stability under fault conditions. The proposed system has effectively maintained the power system stability under fault conditions. The comparative analysis of fuzzy- and firefly-based UPFC system is in Table 1.

drksujatha23@gmail.com

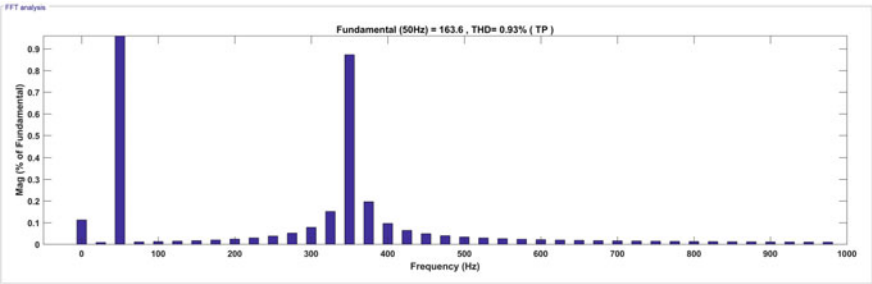


Fig. 11 THD for three-phase fault with UPQC + fuzzy controller

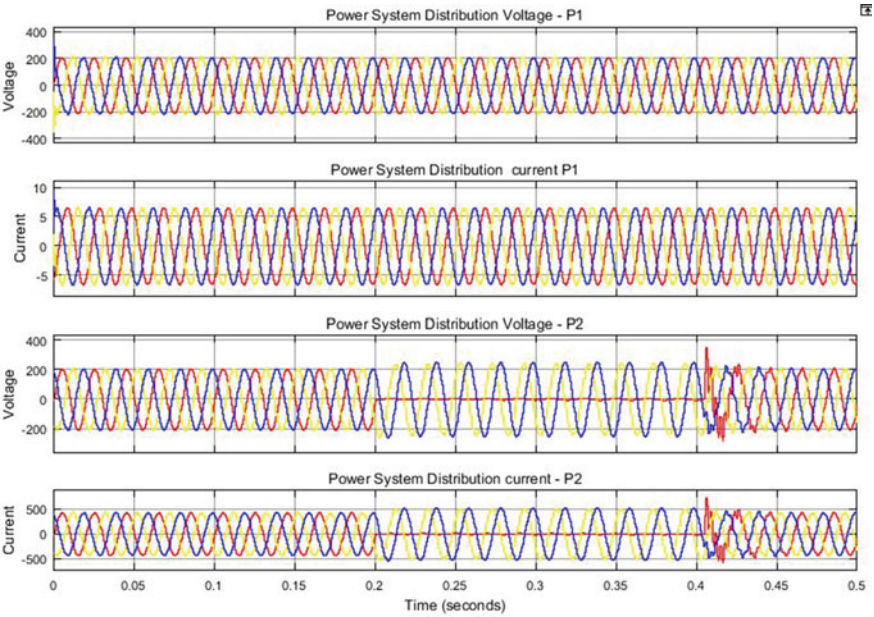


Fig. 12 Line to ground fault with UPQC + fuzzy controller

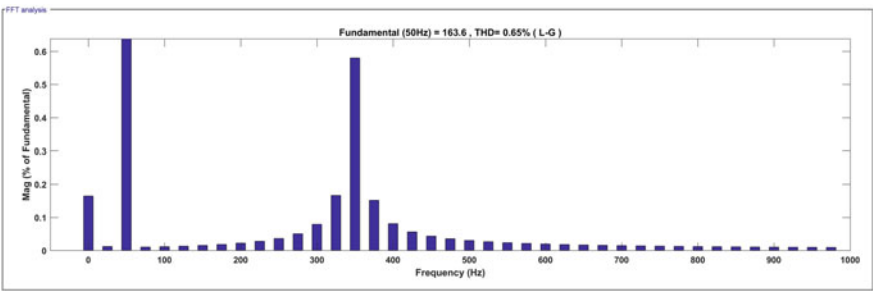


Fig. 13 THD for line to ground fault with UPQC + fuzzy controller

drksujatha23@gmail.com

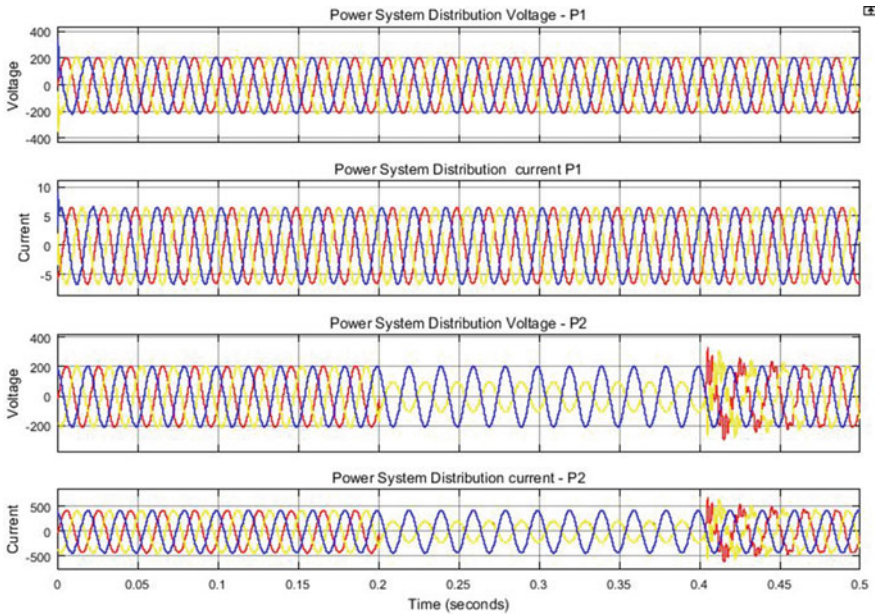


Fig. 14 Line to line fault with UPQC + fuzzy controller

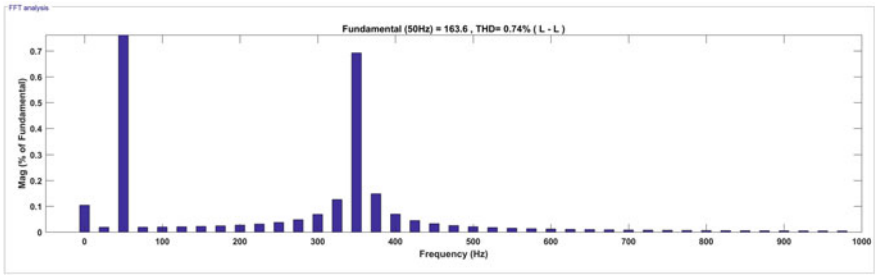


Fig. 15 THD for line to line fault with UPQC + fuzzy controller

drksujatha23@gmail.com

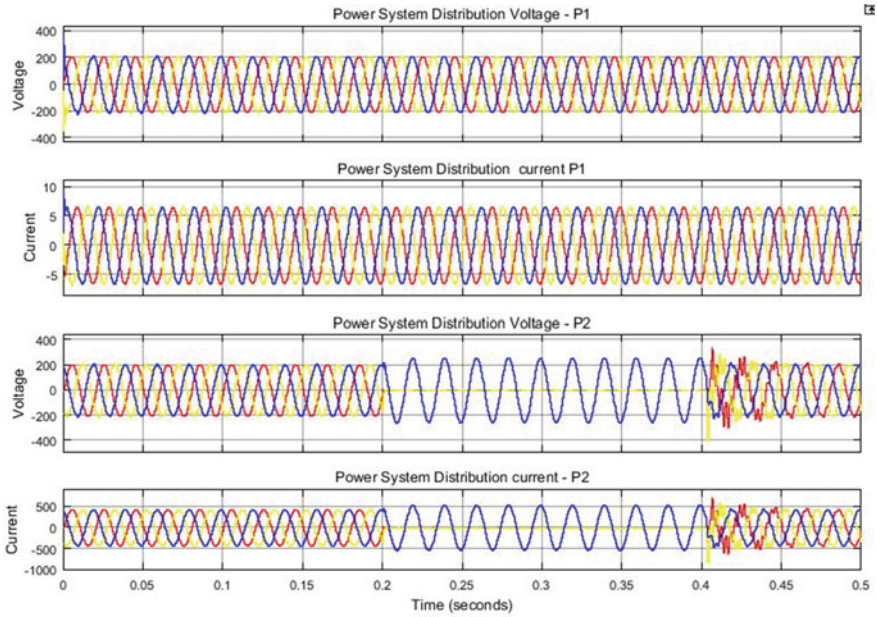


Fig. 16 Double line to ground fault with UPQC + fuzzy controller

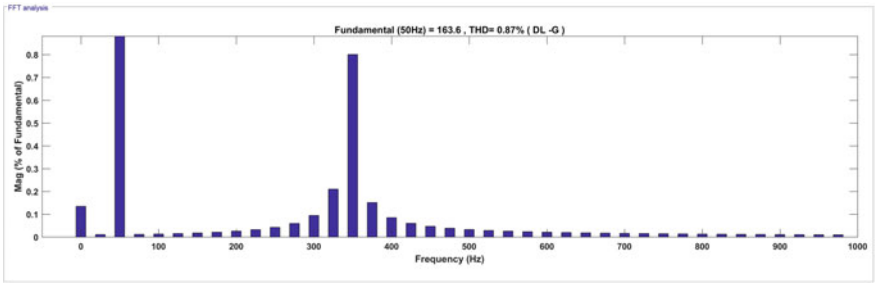


Fig. 17 THD for double line to ground fault with UPQC + fuzzy controller

Table 1 Comparison of THD values for radial distributed voltage with different controller

Faults	Without UPFC (%)	With UPFC and fuzzy (%)	With UPQC and fuzzy (%)
3 phases	37.05	13.7	0.93
L-G	129.1	8.65	0.65
L-L	132.6	12.12	0.74
DL-G	145.5	13.08	0.87

drksujatha23@gmail.com

4 Conclusion

This paper the modeling and analysis of UPQC-based power system under various fault condition such as symmetrical and unsymmetrical faults. The above system is simulated with MATLAB simulation environment with and without UPQC power system at various fault conditions. The design and simulation of using intelligent controller based UPQC power system were analyzed with various fault conditions. The simulated results are evaluated and validated with existing systems as shown in Table 1. Finally, the proposed system has proved the effectiveness of the operation and recommended for voltage improvement radial distribution network.

References

1. Yada HK, Murthy MSR (2016) Operation and control of single-phase UPQC based on SOGI-PLL. In: 2016 7th India international conference on power electronics (IICPE), pp 1–6, Patiala, India
2. Paithankar S, Zende R (2017) Comparison between UPQC, iUPQC and improved iUPQC. In: 2017 third international conference on sensing, signal processing and security (ICSSS), pp 61–64, Chennai
3. Hafezi H, Faranda R (2017) Open UPQC series and shunt units cooperation within Smart LV Grid. In: 2017 6th international conference on clean electrical power (ICCEP), pp 304–310, Santa Margherita Ligure
4. Vadivu US, Keshavan BK (2017) Power quality enhancement of UPQC connected WECS using FFA with RNN. In: 2017 IEEE international conference on environment and electrical engineering and 2017 IEEE industrial and commercial power systems Europe (EEEIC/I&CPS Europe), pp 1–6, Milan
5. Chindris M, Cziker A, Miron A (2017) UPQC—the best solution to improve power quality in low voltage weak distribution networks. In: 2017 international conference on modern power systems (MPS), pp 1–8, Cluj-Napoca
6. Kotturu J, Kumar V, Kothuru S, Agarwal P (2016) Implementation of UPQC for three phase three wire system. In: 2016 IEEE 1st international conference on power electronics, intelligent control and energy systems (ICPEICES), pp 1–6, Delhi
7. Hagh MT, Sabahi M (2016) A single phase unified power quality conditioner (UPQC). In: 2016 IEEE international conference on power system technology (POWERCON), pp 1–4, Wollongong, NSW
8. Abderrahmane AC, Hamid O, Fouad G (2015) Non linear control of the UPQC for grid current and voltage improvement. In: 2015 third world conference on complex systems (WCCS), pp 1–6, Marrakech
9. Hasan M, Ansari AQ, Singh B (2015) Parameters estimation of a series VSC and shunt VSC to design a unified power quality conditioner (UPQC). In: 2015 39th national systems conference (NSC), pp 1–6, Noida
10. Xu Q, Ma F, Luo A, He Z, Xiao H (2016) Analysis and control of M3C-based UPQC for power quality improvement in medium/high-voltage power grid. IEEE Trans Power Electron 31(12):8182–8194
11. Yahiya MAA, Uzair MAR (2016) Performance analysis of DVR, DSTATCOM and UPQC for improving the power quality with various control strategies. In: 2016 biennial international conference on power and energy systems: towards sustainable energy (PESTSE), pp 1–4, Bangalore

drksujatha23@gmail.com

12. Renduchintala UK, Pang C (2016) Neuro-fuzzy based UPQC controller for power quality improvement in micro grid system. In: 2016 IEEE/PES transmission and distribution conference and exposition (T&D), pp 1–5, Dallas, TX
13. Falvo MC, Manganelli M, Faranda R, Hafezi H (2016) Smart n-grid energy management with an open UPQC. In: 2016 IEEE 16th international conference on environment and electrical engineering (EEEIC), pp 1–6, Florence
14. Kotturu J, Agarwal P (2016) Comparative performance analysis of UPQC using two level and three level inverter for three phase three wire system. In: 2016 IEEE 6th international conference on power systems (ICPS), pp 1–6, New Delhi
15. Vani Krishna RS, Mohan P (2016) Design and analysis of UPQC with DG for mitigating power quality issues. In: 2016 international conference on energy efficient technologies for sustainability (ICEETS), pp 101–105, Nagercoil
16. Yasmeena, Das GTR (2016) A review of UPQC topologies for reduced DC link voltage with MATLAB simulation models. In: 2016 international conference on emerging trends in engineering, technology and science (ICETETS), pp 1–7, Pudukkottai

drksujatha23@gmail.com

Design and Comparative Analysis of Various Intelligent Controller Based Efficiency Improvement of Fuel Cell System



M. Venkateshkumar, R. Raghavan, R. Indumathi and Shivashankar Sukumar

Abstract In last decade, the growth of fuel cell power system based research has been reached enormous. A fuel cell's output power depends nonlinearly on the current or voltage due to fuel flow rate, and there exists a unique maximum power point (MPP). Thus, a maximum power point tracking (MPPT) controller is needed to continuously deliver the highest possible power to the load when variations in operation conditions occur. This paper concentrates to analysis and improves the efficiency of PEM fuel cell using various intelligent controllers' techniques based MPPT. The various intelligent controllers have been designed in MATLAB environment and applied PEM fuel Cell power system. The simulation results are evaluated and compared with each other. Finally, the optimum intelligent controller has been chosen based on their performance in improvement of fuel cell efficiency under nonlinear operating conditions.

Keywords Fuel cell · MPPT · Hybrid intelligent · MATLAB

M. Venkateshkumar (✉)
IEEE YP Madras section, Department of EEE, AVIT, Chennai, India
e-mail: venkatmme@gmail.com

R. Raghavan · R. Indumathi
Educational Consultant, Chennai, India
e-mail: egspraghu@yahoo.com

R. Indumathi
e-mail: indhu.success@gmail.com

S. Sukumar
Institute of Power Engineering (IPE), University Tenega Nasional (UNITEN), Kajang, Malaysia
e-mail: shiva.power1985@gmail.com

© Springer Nature Singapore Pte Ltd. 2019

V. S. Rathore et al. (eds.), *Emerging Trends in Expert Applications and Security*,
Advances in Intelligent Systems and Computing 841,
https://doi.org/10.1007/978-981-13-2285-3_33

drksujatha23@gmail.com

1 Introduction

Insufficiency of power generation due to limited availabSility of fossil fuel and lack of cost-effectiveness in rural areas has become a challenge to be solved in order to get better utilization of people's revenue source [1, 2]. Effective solution for this problem is to focus on the use of renewable energy sources during the generation shortage to meet the consumer demand. Because of the concern about the impact of CO_2 and the anxiety on the effect of using fossil fuels based power generation technology on the environment, in recent years, the researchers are focusing on the renewable energy based clean power generation techniques to save the environment from carbon effect. Among the development of renewable energy based power generation techniques, the fuel cell generation have a major role in clean power generation. The above renewable energy source generates electricity using water [3, 4]. The only drawback of these methods of generation is that optimum electricity generation is possible only during optimum fuel flow. In literature survey, authors [5–8] are deliberated about various methodology that has been adopted for minimization of current produced by Fuel Cell (FC), thereby resulting in minimization of fuel consumption in Polymer Electrolyte Membrane (PEM) FC systems. An adaptive technique, namely the Perturb and Observe (P&O), Voltage-based (VMPPT) and current-based (CMPPT) maximum power point tracking controller to minimize the fuel consumption of a fuel cell have been dealt with in detail, has been utilized by the authors.

In this paper, to generate the maximum power from a FC at different fuel flow rates and to safeguard the FC from over-current and voltage collapses across terminals, a MPPT controller for Fuel Cell with fuel flow optimization has been discussed. The detailed study and design of intelligent controllers is in Sect. 2. Finally, Sect. 3 presented the conclusion the improvement of fuel cell efficiency is compared with different hybrid intelligent controllers.

2 Maximum MPPT for Fuel Cell

Maximum Power Point Tracking has become a basic need for renewable energy systems. Tracking maximum power that can be extracted from a renewable energy system will naturally increase the efficiency of the system. In fuel cell power system, the efficiency can be increased up to maximum level by using MPPT controller [9–12]. Maximum power can track by controlling fuel flow rate. There different types of MPPT techniques to implement in a renewable energy system have been discussed.

drksujatha23@gmail.com

2.1 *Fuzzy Logic Controller*

In this method, the fuzzy logic controller is designed for optimum power generation of Fuel Cell system as shown in Figs. 1 and 2. This controller has one input, namely, Fuel Cell Current. Trapezoidal method is used to convert these parameters to fuzzy set [13, 14]. Knowledge-based system has the reference current that is compared with the observed value. Based on the error, IF-THEN rules for selecting fuel flow rate have been made. Finally, the fuzzy set value is converted into a crisp set using the centre of gravity method, and then the signals are fed into a fuel flow operator for control of the fuel flow in fuel cell system at optimum condition and improved efficiency of the system [15–17].

The fuzzy controller has been designed with one input signal and one output signal. Steps to achieve optimum condition are

Step 1: The Fuzzy controller has one input, namely, Fuel cell Current as well as one output signal, namely, pressure of fuel flow (Vide Fig. 3). The fuzzy input signal is converted from crisp set into fuzzy set by using trapezoidal methods and input membership functions are sub divided into three groups, namely, Low, Medium, and High (Vide Fig. 4).

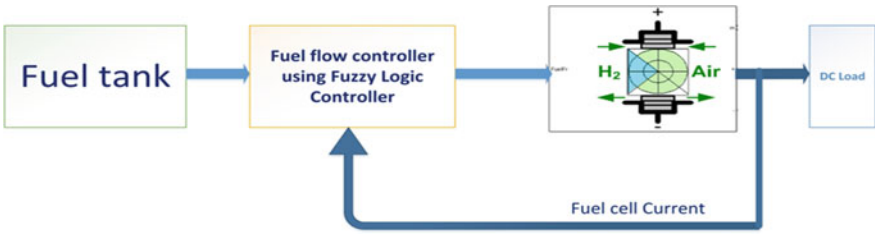


Fig. 1 Fuzzy-based MPPT controller for fuel cell system

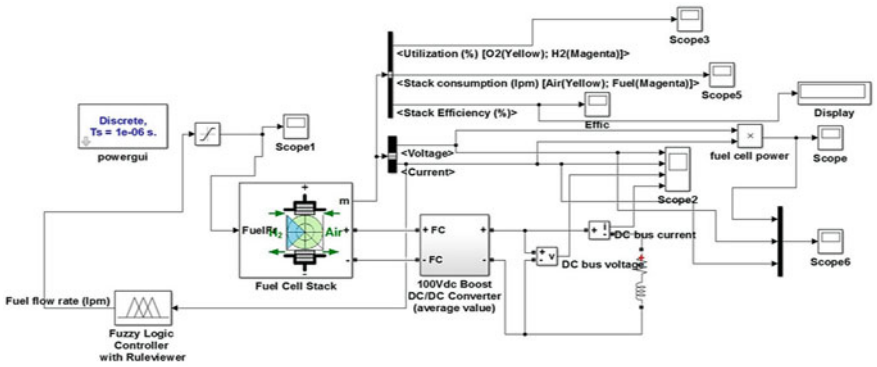


Fig. 2 MATLAB simulation model of fuzzy based MPPT controller for fuel cell system

drksujatha23@gmail.com

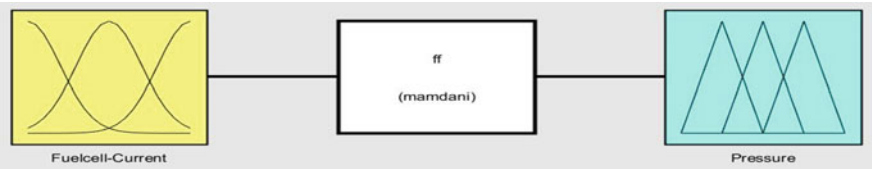


Fig. 3 Fuzzy-based MPPT controller network for fuel cell system

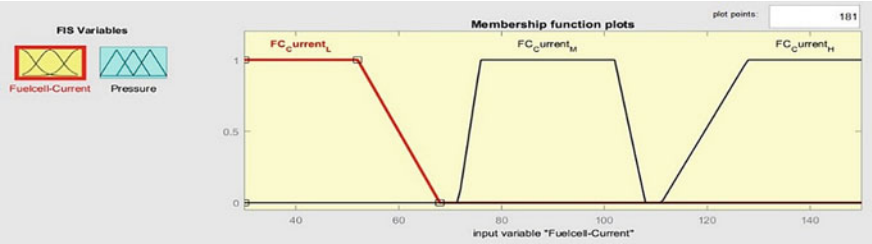


Fig. 4 Fuzzy input membership function for FC current

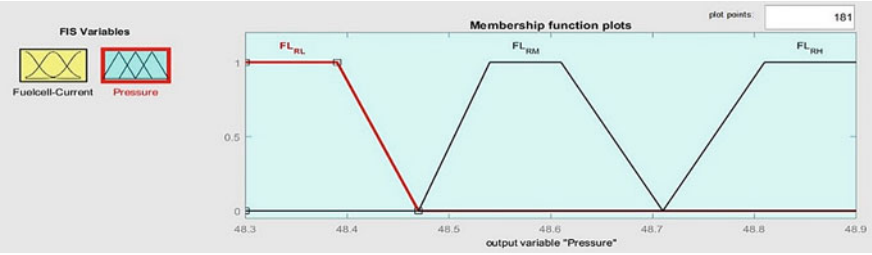


Fig. 5 Fuzzy output membership function for pressure of fuel flow

Step 2: The fuzzy controller has one output signal such as pressure of fuel flow. The fuzzy output signal is converted from crisp set into fuzzy set by using trapezoidal methods (Vide Fig. 5).

Step 3: Fuzzy interference rules are developed by using IF-Then condition after designing input and output membership function of proposed controller (Vide Fig. 6).

Results: The Fuzzy controller has been implemented in Fuel cell MPPT simulation model and simulated in MATLAB environmental. The simulation result of fuel cell efficiency graph is presented in Fig. 7.

drksujatha23@gmail.com

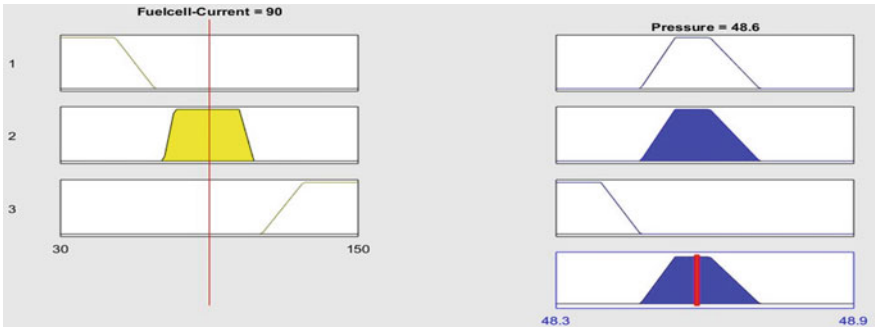


Fig. 6 Fuzzy rules for MPPT of fuel cell system

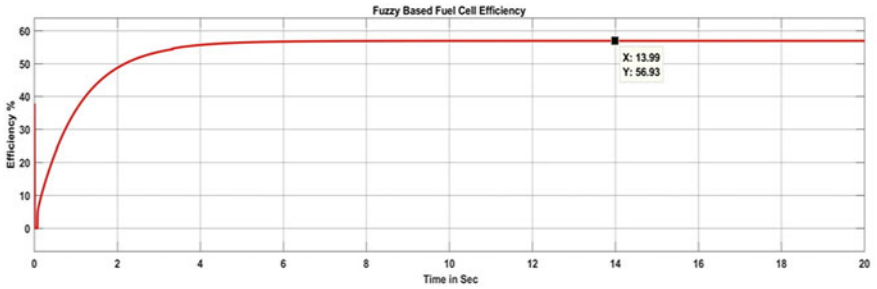


Fig. 7 Fuel cell stack maximum efficiency graph by using fuzzy controller

2.2 ANN Based MPPT of Fuel Cell

In this method, the ANN logic controller is designed for optimum power generation of Fuel Cell system. This controller has one input, namely, Fuel Cell Current (Vide Fig. 8). The input and target data are fed into ANN controller to provide training for the above data using back Propagation algorithm. The trained data are validated and tested to provide best solution. The ANN controller network is developed after completing the processes of training, testing, and validation. Finally ANN controller has been implemented for optimum fuel flow for Fuel Cell system. The ANN controller generates the fuel flow rate based on input data and then the signals is fed into a fuel flow operator for control of the fuel flow in fuel cell system at optimum condition and improved efficiency of the system.

Steps to achieve optimum condition are:

Step 1: The ANN controller, network has been training using input data, such as PV parameters under various weather conditions and target data (duty Cycle) based on changing inputs. The above process has been developed by using ANN training tools in MATLAB environment (Vide Fig. 9).

drksujatha23@gmail.com

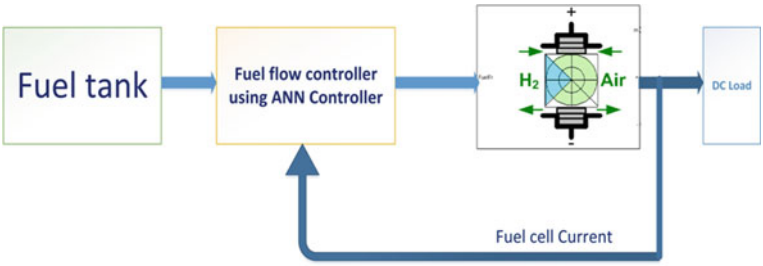


Fig. 8 ANN-based MPPT controller for fuel cell system

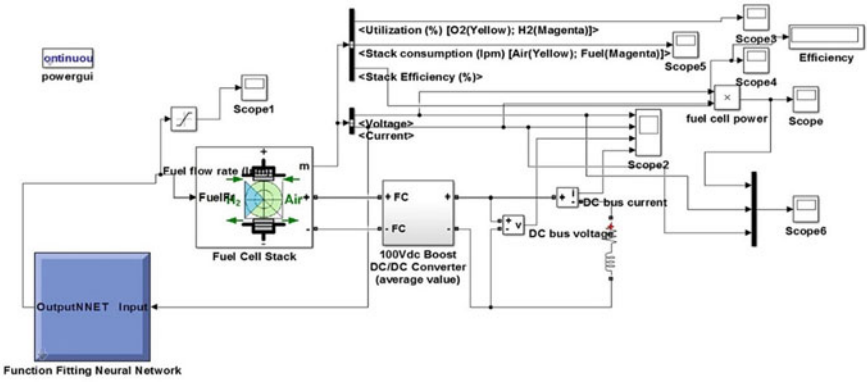


Fig. 9 MATLAB simulation model of ANN based MPPT controller for fuel cell system

Step 2: To analyse and validate checks for training data, inputs and target (Vide Figs. 10 and 11).

Step 3: Again provide training for ANN controller for input data and exact output data analysis and validation checks for training data, inputs and output. Finally evaluate error value in between target and output data (Vide Figs. 12 and 13).

Step 4: To analysis best validation epoch for the trained ANN controller. The controller overall performance is presented in Fig. 14. Finally trained ANN controller has been developed as a simulation model. The ANN controller generates the pressure of fuel flow rate based on input data.

Results: The ANN controller has been implemented in MPPT simulation model and simulated. The simulation result of fuel cell efficiency graph is presented in Fig. 15.

drksujatha23@gmail.com

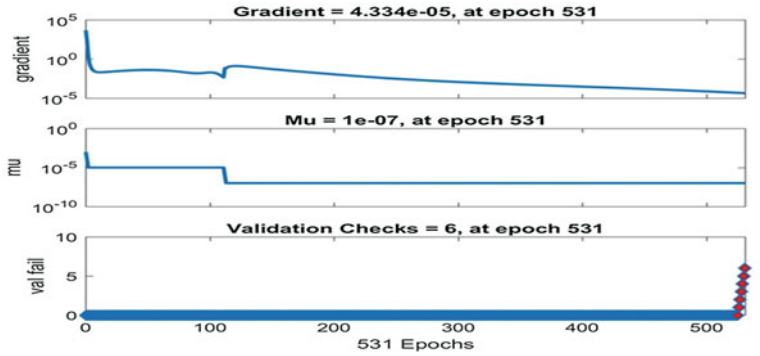


Fig. 10 ANN validation checks and gradient waveform for MPPT of FC

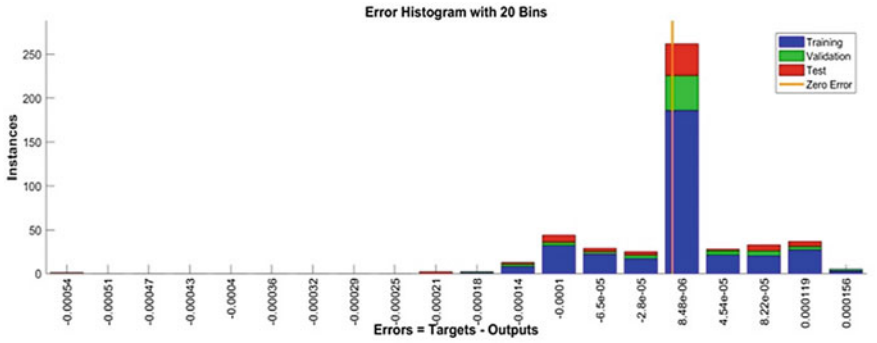


Fig. 11 ANN error waveform after training data for MPPT of FC

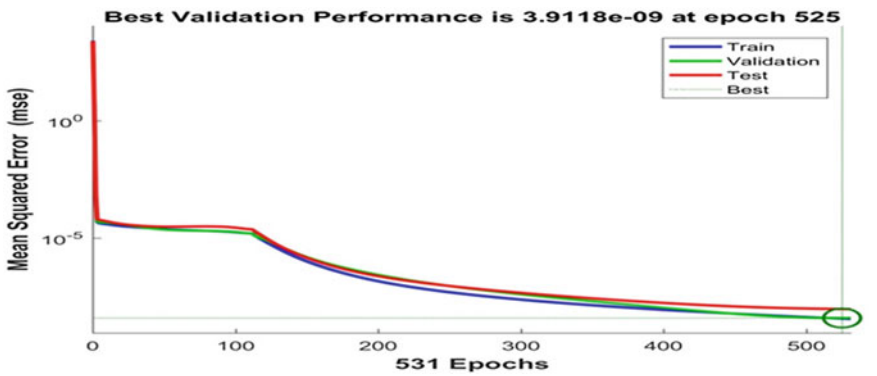


Fig. 12 ANN validation waveform for MPPT of FC

drksujatha23@gmail.com

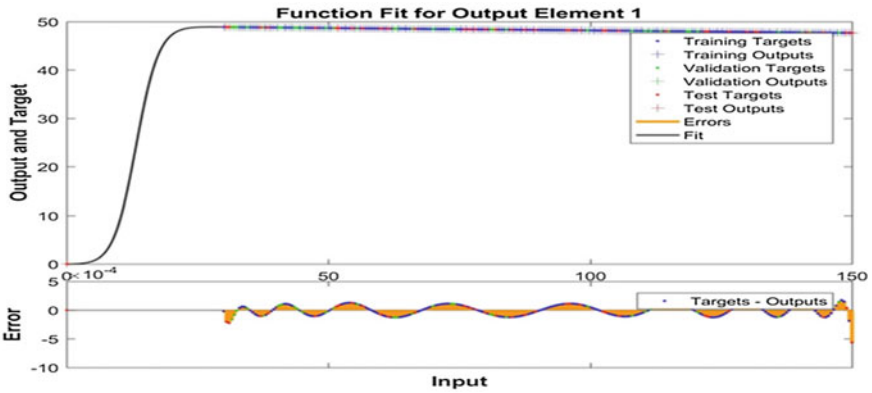


Fig. 13 ANN controller target, output and error waveform for MPPT of FC

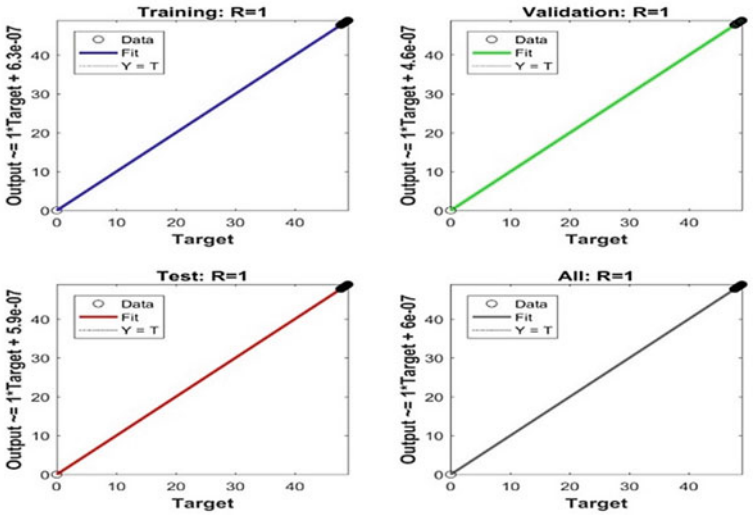


Fig. 14 ANN network training, testing, validation and overall performance waveform for MPPT of FC

2.3 Anfis-Based Mppt

In this method, the ANFIS logic controller is designed for optimum power generation of Fuel Cell system (Vide Fig. 16). This controller has two major parts such as ANN controller design the input and output membership function of fuzzy logic controller and form the fuzzy IF-then rules. Finally, the fuzzy controller generates the optimal pressure of fuel flow based on input data. The above intelligent controller has been simulated in MATLAB environment.

drksujatha23@gmail.com

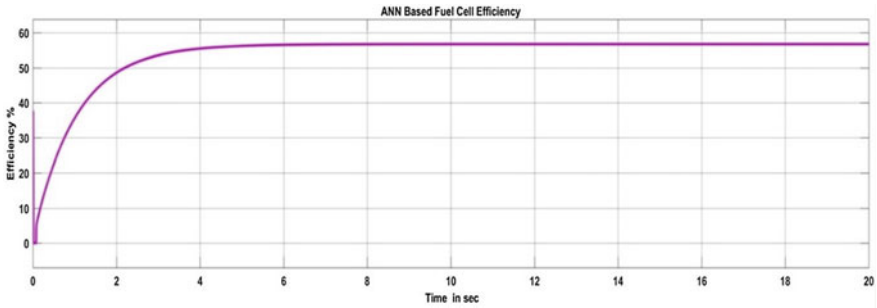


Fig. 15 Fuel cell stack maximum efficiency graph by using ANN controller

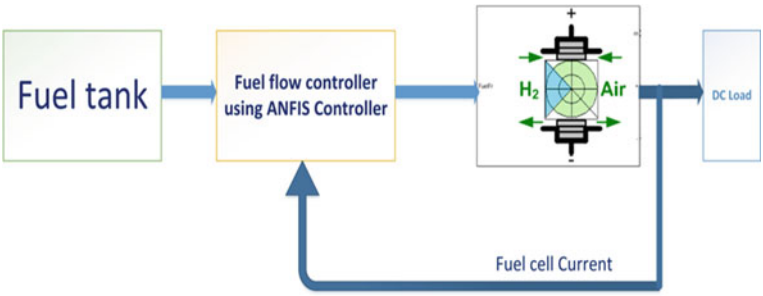


Fig. 16 ANFIS-based MPPT controller for fuel cell system

In this controller, there is one input, namely, Fuel Cell Current. The input and target data are fed into ANN controller to provide training for the above data using back propagation algorithm. The trained data are validated and tested to provide best solution. The ANN controller network develops fuzzy input and output membership functions and then fuzzy IF-Then rules, after completing the processes of training, testing, and validation. Finally, ANFIS controller has been implemented for optimum fuel flow for Fuel Cell system. The ANFIS controller generates the fuel flow rate based on input data and then the signals are fed into a fuel flow operator, for control of the fuel flow in fuel cell system at optimum condition and improved efficiency of the system.

Steps to achieve optimum condition are

Step 1: The ANFIS controller network has been training by using input data namely Fuel cell current under various fuel flow conditions and target data (pressure) based on changing inputs. The above process has been developed by using Neuro-Fuzzy training tools in MATLAB environment (Vide Fig. 17).

Step 2: The trained data are validated and tested for provide best solution. The ANN controller network is developing fuzzy input and output membership functions and

drksujatha23@gmail.com

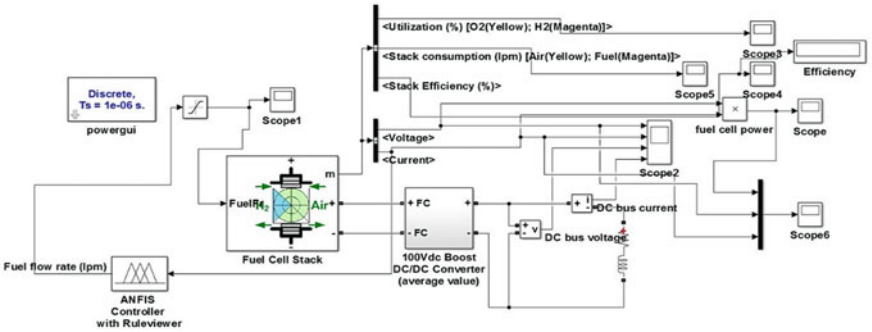


Fig. 17 MATLAB simulation model of ANFIS based MPPT controller for fuel cell system

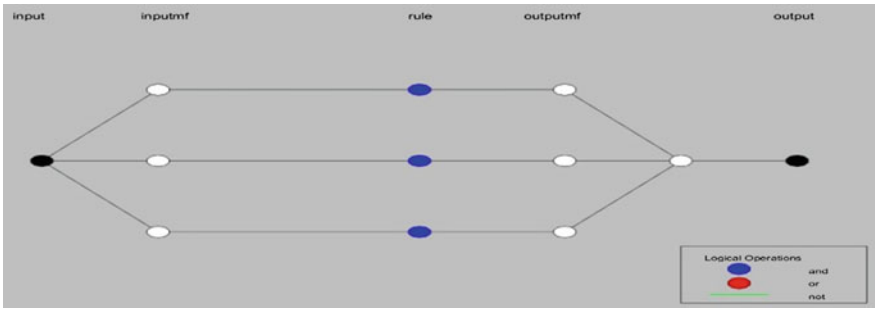


Fig. 18 ANFIS controller network for MPPT of FC system

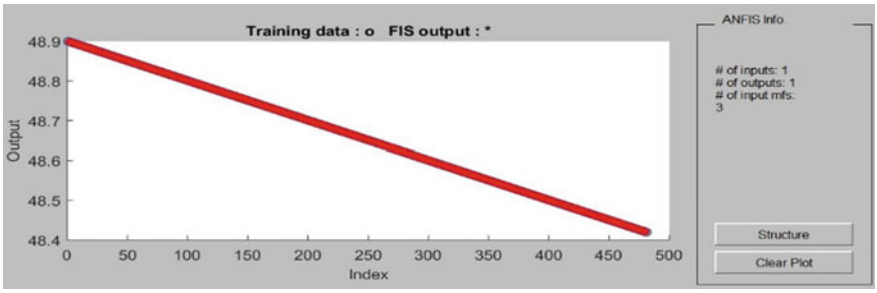


Fig. 19 ANFIS controller training waveform for MPPT of FC system

then fuzzy IF-Then rules after completing the processes of training, testing and validation (Vide Figs. 18, 19, 20).

Step 3: Finally, ANFIS controller has been implemented for MPPT control of PV system. The ANFIS controller generates optimal value of pressure rate for fuel flow based on input data.

drksujatha23@gmail.com

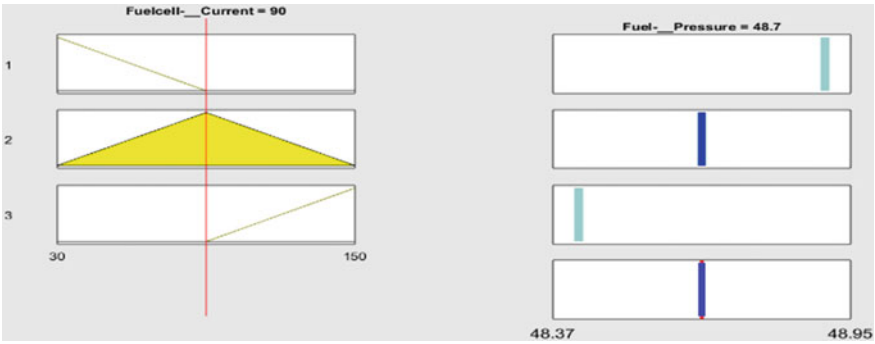


Fig. 20 ANFIS controller rules waveform for MPPT of FC system

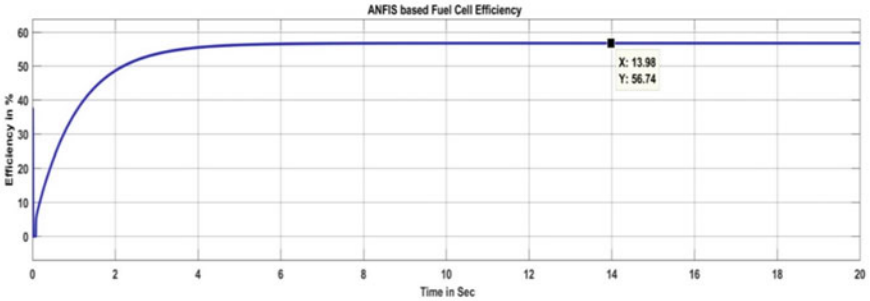


Fig. 21 Fuel cell stack maximum efficiency waveform by using ANFIS controller

Results: The ANFIS controller has been implemented in MPPT simulation model and simulated in MATLAB Simulink environment. The simulation result of fuel cell efficiency graph is presented in Fig. 21.

2.4 Hybrid Fuzzy and Firefly Algorithm

The proposed hybrid fuzzy and firefly has been designed for MPPT controller of fuel cell systems (Vide Fig. 22). The two major intelligent controllers have been designed for optimize the fuel flow rate and achieve the objective of MPPT techniques [18]. The above hybrid intelligent controller has been simulated in MATLAB environment as shown in Fig. 23.

The following steps are used for MPPT controller design

Step 1: The Fuzzy controller has one input namely FC Current as well as one output signals, namely, fuel flow rate in terms of pressure of fuel cell system. The fuzzy input signals are converted from crisp set into fuzzy set by using trapezoidal

drksujatha23@gmail.com

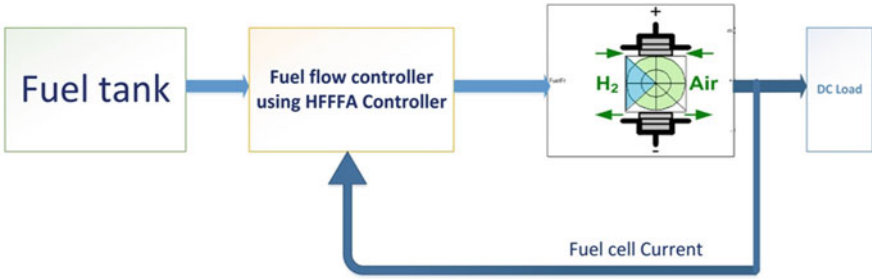


Fig. 22 Hybrid fuzzy and firefly based MPPT controller for fuel cell system

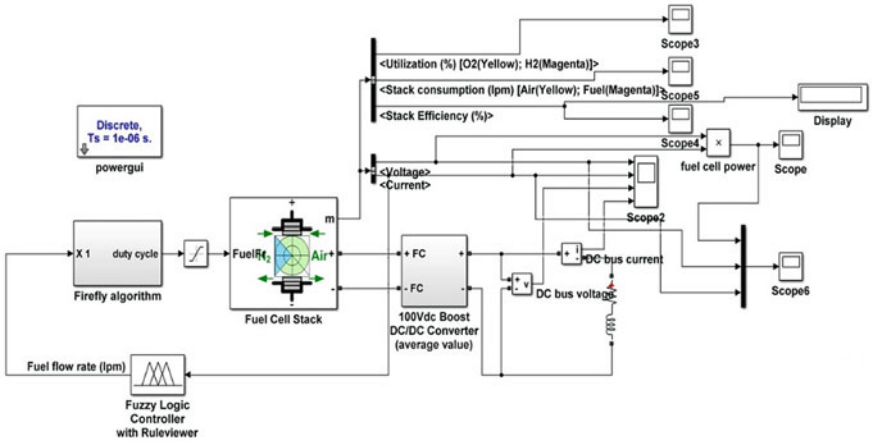


Fig. 23 MATLAB simulation model of hybrid fuzzy and firefly based MPPT controller for fuel cell system

method. Input membership functions are sub-divided into three groups, namely, Low, Medium, and High.

Step 2: The fuzzy controller has one output signal, namely, fuel flow rate in terms of pressure of fuel cell system. The fuzzy output signal is converted from crisp set into fuzzy set by using trapezoidal method.

Step 3: Fuzzy interference rules are developed using IF-Then condition after designing input and output membership function of proposed controller.

Step 4: Defuzzification process has been applied to convert fuzzy set value of fuzzy controller output signals into crisp set value of fuzzy controller output signals by using centroid method. Finally, the fuzzy controller output signal (pressure -Pmax) is fed to firefly controller.

Step 5: Constants of Firefly algorithm, namely βo, γ, n, α, population size N are fixed.

drksujatha23@gmail.com

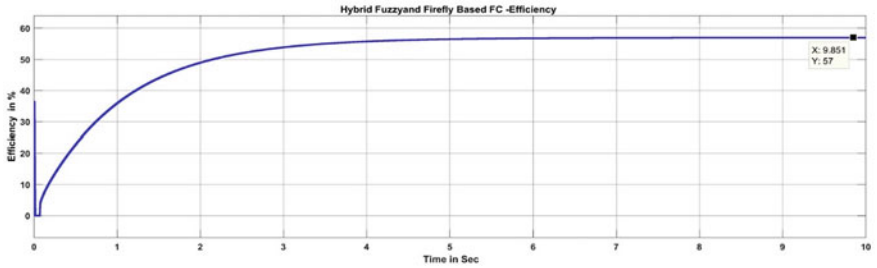


Fig. 24 Fuel cell stack maximum efficiency waveform by using hybrid fuzzy and firefly controller

Where

β_o Initial attractiveness (1.2)
γ light intensity (0.9)
α random movement factor (0.99)

In this algorithm, the position of the firefly is taken as a fuel flow rate p of the fuel cell system.

Step 6: In this step, the fireflies are positioned in allowable solution space between Pressure min to Pressure max. The Pmax value is generated by fuzzy controller. This position of each firefly represents the fuel flow rate of fuel cell system.

Results: The hybrid Fuzzy and Firefly algorithm controller has been implemented in MPPT simulation model and simulated. The simulation result of fuel cell efficiency graph is presented in Fig. 24.

2.5 Hybrid Anfis and Firefly Algorithm

The proposed hybrid ANFIS and firefly has been designed for MPPT controller of fuel cell systems (Vide Fig. 25). The two major intelligent controllers have been designed to optimize the fuel flow rate and achieve the objective of MPPT techniques. The above intelligent controller has been simulated in Matlab environment (Vide Fig. 26).

The following steps are used for MPPT controller design:

Step 1: The ANFIS controller network has been training by using input data namely Fuel Cell Current under various pressures of fuel flow conditions and target data (Pressure of fuel flow) based on changing inputs. The above process has been developed using Neuro-Fuzzy training tools in MATLAB environment.
Step 2: The trained data are validated and tested to provide best solution. The ANN controller network is developing fuzzy input and output membership functions and

drksujatha23@gmail.com

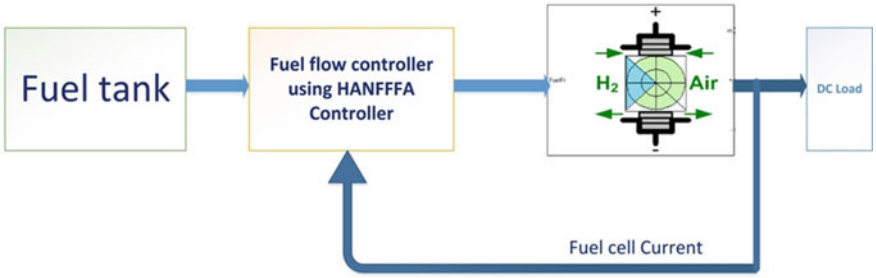


Fig. 25 Hybrid ANFIS and firefly based MPPT controller for fuel cell system

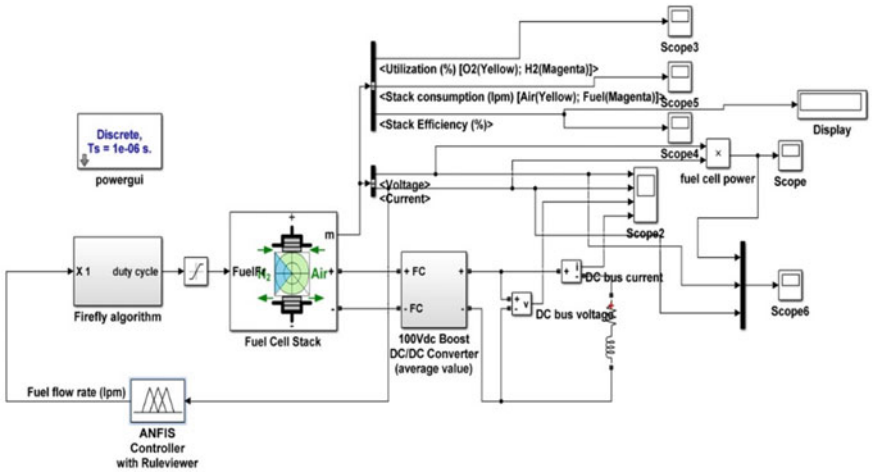


Fig. 26 MATLAB simulation model of hybrid ANFIS and firefly based MPPT controller for fuel cell system

fuzzy IF-Then rules, after completing the processes of training, testing, and validation.

Step 3: Finally, ANFIS controller has been implemented for MPPT control of PV system. The ANN controller generates the duty cycle based on input data and then the signal is fed into a PWM generator to generate the pulse for DC–DC converter.

Step 4: Constants of Firefly algorithm, namely βo, γ, n, α, population size N are fixed

where

β_o Initial attractiveness (1.2)
γ light intensity (0.9)
α random movement factor (0.99)

drksujatha23@gmail.com

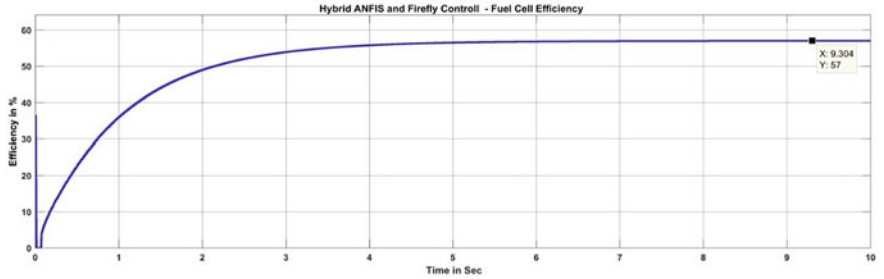


Fig. 27 Fuel cell stack maximum efficiency waveform by using hybrid ANFIS and firefly controller

Table 1 Comparison of fuel cell efficiency with different controllers

Name of the controller	PI in %	Fuzzy in %	ANN in %	ANFIS in %	HFFFA in %	HANFFFA in %
Fuel cell stack efficiency	55.00	56.85	56.74	56.74	57.00 Time taken 9.85 s	57.00 Time taken 9.3 s

In this algorithm, the position of the firefly is taken as a fuel flow rate p of the fuel cell system.

Step 5: In this step, the fireflies are positioned in allowable solution space between Pressure min to Pressure max. The Pmax value is generated by ANFIS controller. This position of each firefly represents the fuel flow rate of fuel cell system.

Results: The hybrid Fuzzy and Firefly algorithm controller has been implemented in MPPT simulation model and simulated. The simulation result of fuel cell efficiency graph is presented in Fig. 27.

In Table 1 analysis of performance of various intelligent controllers while to operating at optimal in efficiency of 6000 W PEM fuel cell stack is tabulated. Based on controller performance the hybrid ANFIS and Firefly controller has been selected for operate optimal efficiency of 6000 W PEM fuel cell stack.

3 Conclusion

In this paper, various hybrid intelligent controllers have been discussed and modelled to improve the fuel cell efficiency. The various intelligent controllers are simulated, and results are analysed under various conditions. Finally, the intelligent controllers' results are evaluated for MPPT techniques of fuel cell system. In Table 1 analysis of performance of various intelligent controllers while to operating

drksujatha23@gmail.com

at optimal in efficiency of 6000 W PEM fuel cell stack is tabulated. Based on controller performance the hybrid ANFIS and Firefly controller has been selected for operate optimal efficiency of 6000 W PEM fuel cell stack.

References

1. Somaiah B, Agarwal V (2016) Distributed maximum power extraction from fuel cell stack arrays using dedicated power converters in series and parallel configuration. IEEE Trans Energy Convers 31(4):1442–1451
2. Wang MH, Huang ML, Jiang WJ, Liou KJ (2016) Maximum power point tracking control method for proton exchange membrane fuel cell. IET Renew Power Gener 10(7):908–915
3. Rezk H (2016) Performance of incremental resistance MPPT based proton exchange membrane fuel cell power system. In: 2016 eighteenth international middle east power systems conference (MEPCON), Cairo, pp 199–205
4. Patil SN, Prasad RC (2015) Design and development of MPPT algorithm for high efficient DC-DC converter for solar energy system connected to grid. In: 2015 international conference on energy systems and applications, Pune, pp 228–233
5. Karami N, Khoury LE, Khoury G, Moubayed N (2014) Comparative study between P&O and incremental conductance for fuel cell MPPT. In: International conference on renewable energies for developing countries 2014, Beirut, pp 17–22
6. Ramos-Paja CA, Spagnuolo G, Petrone G, Giral R, Romero A (2010) Fuel cell MPPT for fuel consumption optimization. In: Proceedings of 2010 IEEE international symposium on circuits and systems, Paris, pp 2199–2202
7. Rezk H (2016) Performance of incremental resistance MPPT based proton exchange membrane fuel cell power system. In: 2016 eighteenth international middle east power systems conference (MEPCON), Cairo, pp 199–205
8. Ettihir K, Boulon L, Agbossou K, Kelouwani S (2012) MPPT control strategy on PEM fuel cell low speed vehicle. In: 2012 IEEE vehicle power and propulsion conference, Seoul, pp 926–931
9. Karami N, Outbib R, Moubayed N (2012) Fuel flow control of a PEM fuel cell with MPPT. In: 2012 IEEE international symposium on intelligent control, Dubrovnik, pp 289–294
10. Pachauri RK, Chauhan YK (2014) Hydrogen generation/pressure enhancement using FC and ANN based MPPT assisted PV system. In: 2014 innovative applications of computational intelligence on power, energy and controls with their impact on humanity (CIPECH), Ghaziabad, pp 427–432
11. Mane S, Kadam P, Lahoti G, Kazi F, Singh NM (2016) Optimal load balancing strategy for hybrid energy management system in DC microgrid with PV, fuel cell and battery storage. In: 2016 IEEE international conference on renewable energy research and applications (ICRERA), Birmingham, pp 851–856
12. Liu J, Zhao T, Chen Y (2017) Maximum power point tracking with fractional order high pass filter for proton exchange membrane fuel cell. IEEE/CAA J Automatica Sinica 4(1):70–79
13. Gördesel M, Canan B, Günlü G, Sanlı AE (2015) Fuel cell powered hybrid system controlled by the maximum peak power tracking technique. In: 2015 twelve international conference on electronics computer and computation (ICECCO), Almaty, pp 1–3
14. Djoudi H, Badji A, Benyahia N, Zaouia M, Denoun H, Benamrouche N (2015) Modeling and power management control of the photovoltaic and fuel cell/electrolyzer system for stand-alone applications. In: 2015 4th international conference on electrical engineering (ICEE), Boumerdes, pp 1–6

drksujatha23@gmail.com

15. Venkateshkumar M, Sarathkumar G, Britto S (2013) Intelligent control based MPPT method for fuel cell power system. In: 2013 international conference on renewable energy and sustainable energy (ICRESE), Coimbatore, pp 253–257
16. Sarvi M, Barati MM (2010) Voltage and current based MPPT of fuel cells under variable temperature conditions. In: 45th international universities power engineering conference UPEC2010, Cardiff, Wales, pp 1–4
17. Ramos-Paja CA, Spagnuolo G, Petrone G, Giral R, Romero A (2010) Fuel cell MPPT for fuel consumption optimization. In: Proceedings of 2010 IEEE international symposium on circuits and systems, Paris, pp 2199–2202
18. Venkateshkumar M, Indumathi R (2017) Comparative analysis of hybrid intelligent controller based MPPT of fuel cell power system. In: 2017 IEEE international conference on smart technologies and management for computing, communication, controls, energy and materials (ICSTM), Chennai, pp 155–159

drksujatha23@gmail.com

Analysis of Load Balancing Algorithms Using Cloud Analyst



Jyoti Rathore, Bright Keswani and Vijay Singh Rathore

Abstract Cloud computing delivers the services of cloud efficiently and effectively on pay-per usage basis to consumers. As the number of consumers and requests for the services are increasing day by day in cloud computing need of load balancing occurs. For efficient and effective management and usage of cloud service provider's resources, already many load balancing algorithms have been proposed. In this paper comparative analysis of existing algorithms has been performed using simulator, i.e., Cloud Analyst. Cloud Analyst is a GUI-based toolkit that performs testing and simulation. In this paper existing Throttled, Round Robin, ESCE, FCFS, and SJF are compared. Cloud Analyst simulation results shows significant outcomes in terms of data center processing time, total cost, and response time in cloud computing environment.

Keywords Cloud computing · Load balancing · Load balancing algorithms
Cloud analyst · Virtual machine · Simulator · Overall response time
Data center processing time, etc.

J. Rathore (✉)
Suresh Gyan Vihar University, Jaipur, India
e-mail: jyoti.rathore131@gmail.com

B. Keswani
Department of Computer Applications, Suresh Gyan Vihar University, Jaipur, India
e-mail: kbright@rediffmail.com

V. S. Rathore
Department of Computer Science & Engineering, Jaipur Engineering College & Research Centre, Jaipur, Rajasthan, India
e-mail: vijaydiamond@gmail.com

© Springer Nature Singapore Pte Ltd. 2019

V. S. Rathore et al. (eds.), *Emerging Trends in Expert Applications and Security*,
Advances in Intelligent Systems and Computing 841,
https://doi.org/10.1007/978-981-13-2285-3_34

drksujatha23@gmail.com

1 Introduction

In Cloud computing users have on-demand and convenient access of shared pool of computing resource such as application and services, storage, network, etc., on pay-per use basis [1]. In a distributed environment clients request are randomly generated in any processor and because of that load imbalance occurs load balancing is done [2]. The load balancing transfers the excess load from overloaded processor to under-loaded processor [3]. This paper focuses on comparative analysis of existing load balancing algorithms on data center processing time, total cost and response time parameters using cloud analyst.

2 Load Balancing in Cloud Environment

Consumer satisfaction for using cloud services is achieved through Load Balancing, which effectively utilize resource and improve response time of the jobs [4]. Load Balancing algorithm decides which VM is to allocate when request is made by cloud consumer [5]. Some VM load balancing algorithms are:

Round Robin Load Balancing Algorithm
It is a static approach [6]. It places the newly coming cloudlets on the available virtual machines (VM) in a circular manner. The first VM is chosen randomly. Its advantages are its simplicity and easy implementation [4]. Its drawbacks are load imbalance and in case VM is not free, then cloudlets wait in the waiting queue [7].

Throttled Load Balancing Algorithm
It is a dynamic approach, Users request are submitted to the Data Center Controller (DCC). VMLoadBalancer (VMLB) maintains VM list and their states. First all VMs states are set available. DCC asks VMLB about VM. VMLB checks the index table and return VM id which can handle particular load to DCC. DCC allots the requests to that particular VM [7].

Equally Spread Current Execution (ESCE) load Balancing Algorithm
ESCE also called Active VM Load Balancing algorithm. It equally distributes the workload on each VM in data center. VMLB maintains VMs list along with the number of request already allotted to that particular VM. DCC asks VMLB about VM allocation [8]. VMLB send id of VM that can handle load to the DCC. DCC allots requests to that VM. VMLB inspect the overloaded VMs. If found overloaded, then VMLB moves some load to an idle or an under-loaded VM [9].

First Come First Serve (FCFS) load Balancing Algorithm
First the client requests are assigned to the job Queue. The first request which comes in the queue will get executed first. It is not a preemptive disciple. The drawback of this algorithm is that long task gets executed and small task waits or important task waits and unimportant task get executed [2].

drksujatha23@gmail.com

SJF load Balancing Algorithm
SJF stands for "Shortest Job First". This algorithm depends on the CPU burst time of each task. When CPU is available the task with smallest CPU burst time gets executed. If two tasks with same CPU burst time are present then first-come-first-serve algorithm is used.

3 Experimental Setup

In this study, the proposed VM Load Balancing algorithm is implemented using the following software's and tools: Windows7 Operating System, Eclipse Neon.3, JDK 1.8 and Cloud Analyst tool. This algorithm is implemented for IaaS (Infrastructures as a Service) model in a simulated cloud environment.

Cloud Analyst: It is a GUI-based toolkit that performs testing and simulation. Cloud Analyst is an extension to cloudsim simulator. It helps researcher to focus on the different parameters used for simulation rather instead of programming details [10]. Key elements of Cloud Analyst simulator are:

Region: Cloud Analyst toolkit split the World into six regions; these regions are six continents, i.e., Australia, South America, North America, Asia, Africa, and Europe. The essential components of cloud analyst User base and datacenter resides in these regions.

UserBase: A group of users are considered as a single unit in Cloud Analyst which involved in the simulation and call as UserBase. The main task of this component is generation of traffic.

Data Center Controller (DCC): A single cloudsim maps a single DCC. DCC controls all the activities of datacenter like cloudlet request routing, creation and destruction of VMs, etc.

VM Load Balancer: DCC makes communication with VMLB for VMs. VMLB is used to decide the assignment of cloudlets on VMs. Today Cloud Analyst has three VMLB: Round Robin, Throttled, Active Monitoring Load Balancer.

Cloud Application Service Broker: A Service Broker policy manages traffic between data centers and user bases. Cloud Analyst simulator has three service broker policies, i.e., optimize response time, closest data center and dynamically reconfigured.

Simulation Parameters: In addition to existing load balancing policies two new policies are also added to cloud analyst tool named as FSFC and SJF as shown in Fig. 1. For simulation, six user bases named as UB1, UB2, UB3, UB4, UB5, UB6 in region R0, R1, R2, R3, R4, and R5, respectively and four data centers named as D1, D2, D3, D4 in region R0, R4, R2, and R3 respectively as shown in Fig. 2.

drksujatha23@gmail.com

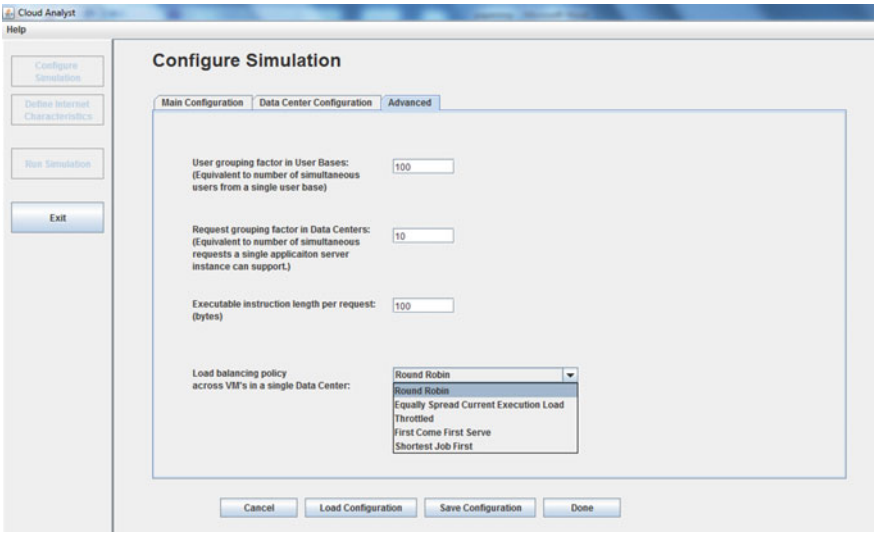


Fig. 1 Two new policies "Shortest Job First" and "First Come First Serve" are added

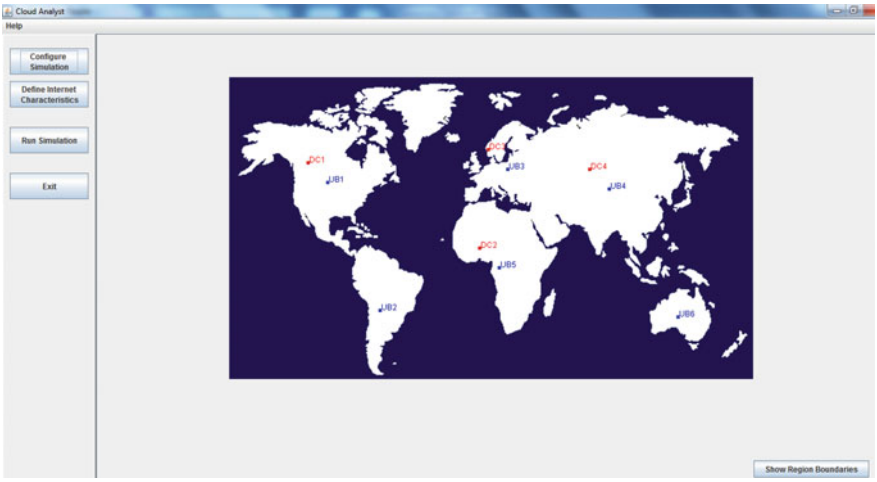


Fig. 2 Illustration of different regions and data centers across the globe

User base configurations such as their requests per user per hr, regions, average peak users, data size of each request, and so on, are shown in Fig. 3. Each physical machine is consolidated with 10 virtual machines as shown in Fig. 3. Each data center is constructed with 15 physical machines, each have following configuration as shown in Fig. 4.

drksujatha23@gmail.com

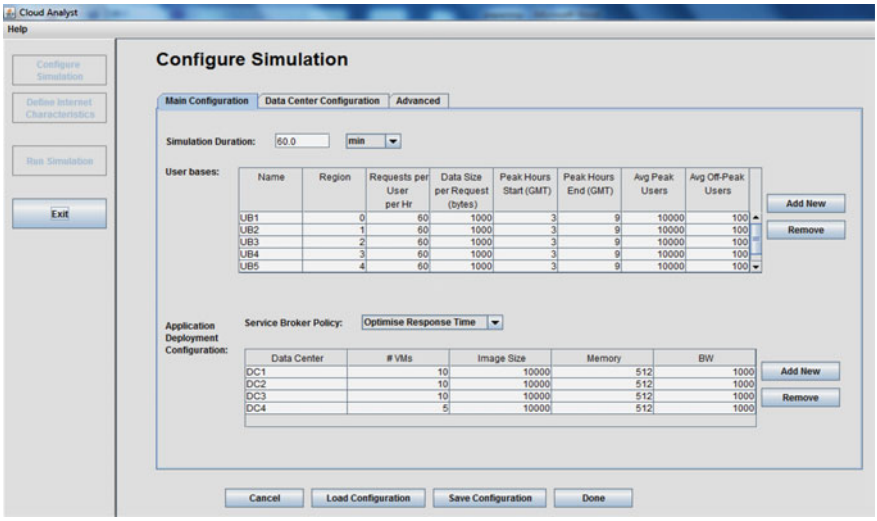


Fig. 3 Configuration of user bases and application deployment configuration with service broker policy and VM parameters

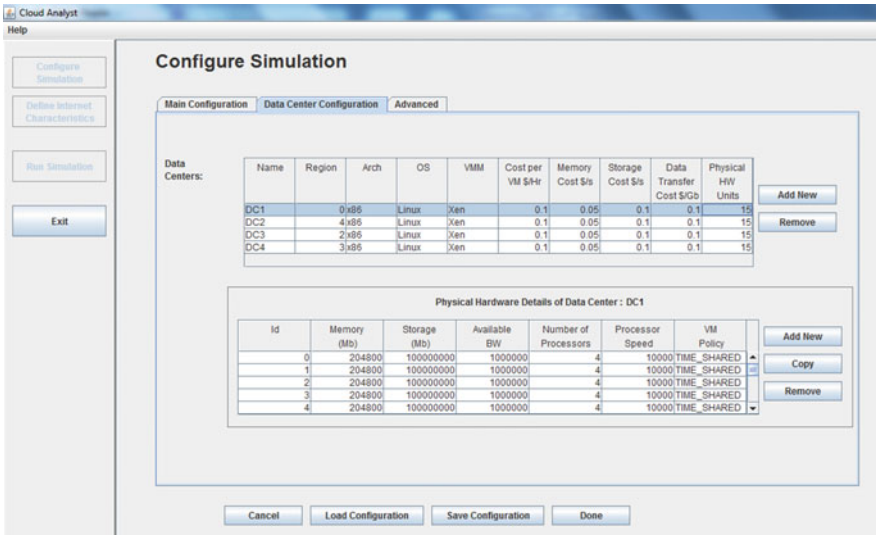


Fig. 4 Configuration of data centers and configuration of physical hardware machine

drksujatha23@gmail.com

4 Performance Analysis

The parameters defined above have been used for each VM load balancing scheduling algorithms one by one. Results are calculated for following matrices: average response time by region, overall response time, request servicing time of data centers, overall processing time of data centers and total cost as shown in Tables 1, 2, 3, 4 and 5 respectively (Figs. 5, 6, 7).

Table 1 Comparison among all load balancing algorithms (average response time (ms))

User bases	Throttled	FCFS	ESCE	Round Robin	SJF
UB1	49.798	50.952	49.716	49.851	50.065
UB2	198.631	198.609	198.579	198.694	198.546
UB3	50.692	49.937	50.594	50.638	50.581
UB4	50.381	50.911	50.384	50.446	50.255
UB5	50.625	49.937	50.733	50.679	50.675
UB6	201.674	201.372	201.728	201.245	201.511

Table 2 Comparison between request servicing time (ms) of data centers (among all algorithms)

Data	Throttled	FCFS	ESCE	Round Robin	SJF
DC1	0.276	0.517	0.246	0.277	0.297
DC2	0.645	0.616	0.654	0.648	0.652
DC3	0.452	0.445	0.451	0.457	0.462
DC4	0.557	0.535	0.579	0.587	0.548

Table 3 Overall response time for all algorithms

Load balancing policies	Throttled	FCFS	ESCE	Round Robin	SJF
Overall response time (ms)	100.67	**70.89**	100.66	100.63	100.64

Note Bold values indicate that the FCFS algorithm provides minimum response time

Table 4 Overall data center processing time for algorithms

Load balancing policies	Round Robin	Throttled	ESCE	FCFS	SJF
Overall data center processing time (ms)	0.42	0.42	**0.41**	0.52	0.43

Note Bold values indicate that the ESCE approach has least overall data center processing time

Table 5 Total cost for algorithms

Load balancing policies	Round Robin	Throttled	ESCE	FCFS	SJF
Total cost	7.05	7.05	**6.52**	12.30	7.05

Note Bold values indicate that the ESCE approach has least Total cost

drksujatha23@gmail.com

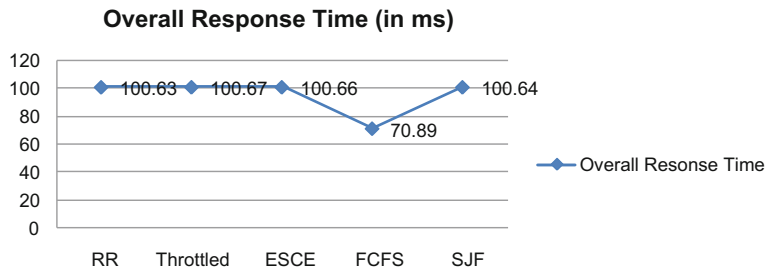


Fig. 5 Overall response time of all algorithms

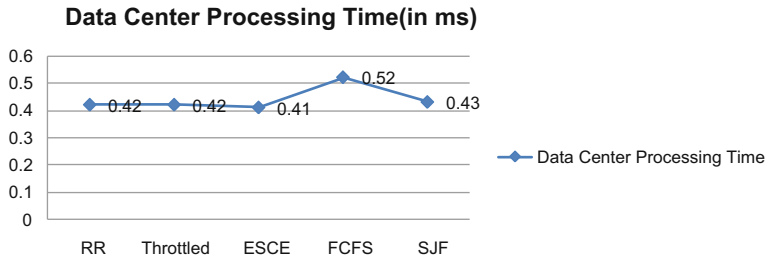


Fig. 6 Data center processing time of load balancing algorithm

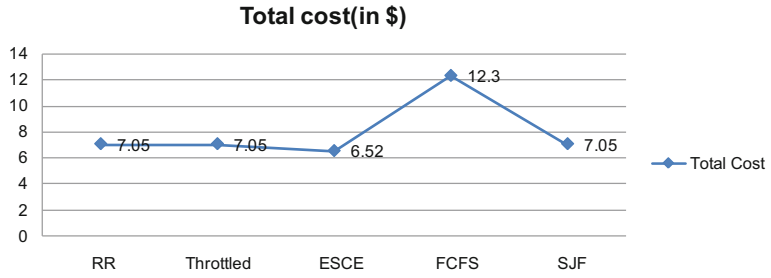


Fig. 7 Total cost of load balancing algorithm

5 Conclusion and Future Scope

By comparative analysis of existing load balancing algorithms it is concluded that when comparison is done on the basis of overall response time parameter, then FCFS load balancing algorithm performs best while comparison is done on the basis of data center processing time and total cost ESCE load balancing algorithm performs best. This work can be extended by using different service broker policies and then analyzing the results of various load balancing techniques.

drksujatha23@gmail.com

References

1. Shakir M, Razzaque A (2017) Performance comparison of load balancing algorithms using cloud analyst in cloud computing. IEEE, pp 509–513
2. Bala K et al (2014) A review of the load balancing techniques at cloud server. Int J Adv Comput Sci Commun Eng 2(I): 6–11
3. Santra S, Mali K (2015) A new approach to survey on load balancing in VM in cloud computing: using CloudSim. In: IEEE international conference on computer, communication and control (IC4-2015)
4. Zegrari F et al (2016) Resource allocation with efficient load balancing in cloud environment. ACM
5. Gabi D et al (2015) Systematic review on existing load balancing techniques in cloud computing. Int J Comput Appl 125(9):16–24
6. Ramesh B (2014) Load balancing in cloud computing—an analysis. In: International conference on security and authentication—SAPIENCE141, pp 125–131
7. Khanchi M, Tyagi S (2016) An efficient algorithm for load balancing in cloud computing. Int J Eng Sci Res Technol. ISSN:2277-9655, Impact Factor: 4.116
8. Kumar R, Prashar T (2015) Performance analysis of load balancing algorithms in cloud computing. Int J Comput Appl 120(7):19–27
9. Behal V, Kumar A (2014) Comparative study of load balancing algorithms in cloud environment using cloud analyst. Int J Comput Appl 97(1201):36–40
10. Singh S, Sharma A (2016) Analysis of load balancing algorithms using cloud analyst. Int J Grid Distrib Comput 9(9):11–24

drksujatha23@gmail.com

Terrain Attribute Prediction Modelling for Southern Gujarat: A Geo-spatial Perspective



Jaishree Tailor and Kalpesh Lad

Abstract Geographical Information Systems (GIS) are crucial to every domain application especially for natural resource, land assessment, and management. Agriculture and terrain conditions are associated with each other where quality and usage of one defines the impact on the other. There is always a concern about terrain related issues and its corresponding native geo-spatial solutions. The present paper focuses upon Terrain Attribute Prediction Modelling for Southern Gujarat. It describes the authors' approach towards determination of the variogram model and its parameters through extensive calibration and validation of experiment. The proposed Terrain Attribute Prediction Model achieves an accuracy of more than 74%.

1 Introduction

The Terrain degradation is a worldwide environmental problem mainly occurring in arid and semiarid regions causing soil degradation. Due urbanization, accumulation of natural, and artificial pollutant on the soil surface makes it either saline, alkaline, or acidic. The geo-spatial variability of such a surface over the landscape is highly sensitive and controlled by a variety of factors. This includes soil parent material, absorbency, water table depth, groundwater quality, and topography. Moreover, management factors like irrigation and drainage as well as climatic factors like rainfall and humidity do contribute towards this situation. Agro stakeholders use a spectrometer to characterize soil salinity, alkalinity, or acidity. They measure the relation of pH with electric conductivity (EC) and organic carbon (OC). This conventional technique lacks visualized characterization through detailed maps due

J. Tailor (✉) · K. Lad
Shrimad Rajachandra Institute of Management and Computer Applications,
Uka Tarsadia University, Bardoli, Surat, India
e-mail: jaishree.tailor@utu.ac.in

K. Lad
e-mail: kalpesh.lad@utu.ac.in

© Springer Nature Singapore Pte Ltd. 2019

V. S. Rathore et al. (eds.), *Emerging Trends in Expert Applications and Security*,
Advances in Intelligent Systems and Computing 841,
https://doi.org/10.1007/978-981-13-2285-3_35

drksujatha23@gmail.com

to increased density of soil samples that makes the process of mapping extensively time consuming and expensive [1, 2]. The above-mentioned issues thus defend the rational behind this work. It also justifies fundamental objective of exploring terrain conditions and its suitability for those villages with for which samples are not collected. This process of exploration involves predictions of soil attributes for this region [3–5]. Consequently, to accomplish this objective especially for the South Gujarat region, model development thus forms an essential aspect the current study [6, 7]. This paper deliberates the sampling strategy as well as determination the geo-spatial variability with extensive experimentation and added different methodologies at every milestone of model development [8, 9].

2 Flow for Terrain Attribute Prediction

Development of predictive models guides in decision-making. With a predictive model, the principal focus is no longer on the data but on the type of theory related to the real world. Here in this case the focus is on prediction of terrain condition for unvisited locations. The process of developing geo-spatial variability estimation models for terrain attributes or any other models commonly known as "calibration" [10, 11]. Given the basic form of a terrain-forecasting model, such as a kriging, calibration involves estimating the values of various constants and parameters in the model structure. Thus, model development effort is also termed as "estimation" and thus requires extensive data and appropriate sampling design [12, 13].

2.1 Sampling and Experimental Design

The authors initiated the process of sampling for the purpose of experimentation. For sampling, the entire dataset composed of 467 villages. The authors divided the data set into an approximate ratio of 2:1 for calibration and validation respectively. For calibration and validation purpose, they used 300 and 167 villages respectively.

After carrying out sampling, the next task the authors performed is construction of experimental design for prediction, three-terrain attributes [14]. The terrain attribute modelling consists of prediction process through Ordinary Kriging that reflects in Fig. 1. The process starts with selection of a control point known as Prediction Point, i.e., terrain attribute prediction for a selected village. The Predictor Points are those villages that used in the process of Variogram calculation and distance. For this, the authors prepared test cases based on sample size to predict a single village from a group of 6–25 villages. Thus to predict the nutrient value of a village, from a sample of 10 villages, the system is designed to randomly select 11 villages. From these 11 villages system uses last 10 villages as predictors while the first village for prediction. Therefore, writers designed 2000 test cases. For each 2000 cases, the next step is to fit the Variogram model. For this work, the models

drksujatha23@gmail.com

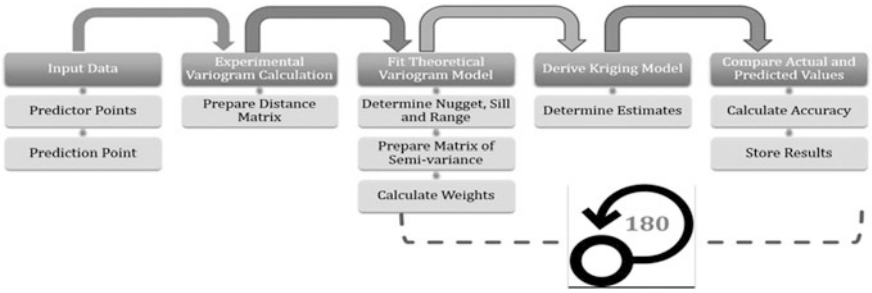


Fig. 1 Terrain attribute flow modelling for calibration

Table 1 NSR alternatives for experimentation

Models	Nuggets	Sills	Ranges
M1 = Gaussian	N1 = Min(ϒ)	S1 = Mean(Max(ϒ), Median(ϒ))	R1 = 0.1 * Diagonal
M2 = Spherical	N2 = Mean (ϒ)/4	S2 = Max(ϒ)	R2 = Max(Dist)
M3 = Exponential	N3 = Zero	S3 = Mean(ϒ)	R3 = Mean(Dist)
		S4 = Max(ϒ) * 0.9	R4 = Max(Dist)/2
			R5 = Max(Dist)/3

utilized are Gaussian, Spherical, and Exponential. Each model covers Nugget (N), Sill (S) and Range (R) parameters. Table 1 presents the NSR with its alternatives derived from the literature. Subsequently computation of semi-variance matrix occurs, followed by weight calculation and kriging estimate of terrain attributes. Further the system compares actual terrain attribute values with predicted and then computation of prediction accuracy and finally result preparation. This process executes for 180 times considering models and its NSR combinations. In all the system calculates 360,000 predictions each attribute [15]. The authors then analyzed the entire predicted data for determination of variogram and its parameter for the south Gujarat region. Finally, through validation phase the authors determined the validity of the model [16].

3 Geo-spatial Variability Modelling

This section describes the maximum likelihood of variogram model and its parameters along with the adopted two approaches.

drksujatha23@gmail.com

3.1 Approaches to Variogram Model Selection

The first approach the authors followed is top most prediction accuracy. As mentioned in previous sub-section out of 360,000 predictions the authors analyzed only those predictions that resulted highest prediction percentage accuracy. For each 180 predictions within a sub-sample, the authors retrieved the highest calculated prediction accuracy. This process continued for all the sub-samples that resulted into 2000 predictions that is around 0.55% of 360,000 predictions.

For analyzing the predicted attributes, the authors adopted the data table format, the first column of which embodies six classes with a class interval of ten representing the prediction accuracy in terms of percentage. The second, third, and fourth column of the data table comprise of percentage participation of each model under each prediction class. The summary of pH appears in Table 2. The incidence of Gaussian model from 2000 (highest accuracy) predictions for pH is around 57.75% over 28.90% and 13.35% for Spherical and Exponential respectively. Likewise, for EC the model frequency is 55.40% that is higher than 30.90% and 13.70% for Spherical and Exponential respectively. Observing the results for OC the Gaussian model significantly dominates its occurrences over the other two models with a score of 54.30% being comparatively higher than 28.70% for Spherical model and 17.00% for Exponential model. The contribution of the Gaussian model in terms of highest prediction accuracy is significantly higher as compared to frequency of occurrence of Spherical and Exponential models. The authors observed a stability at 57.65% that indicates towards success rate of the model giving prediction accuracy higher than 70%. The inference drawn out of first approach acted as prior input for analyzing model determination using the second approach that focuses upon highest occurrence of prediction accuracy with a given model. Therefore, for each terrain attribute, the authors divided 360,000 predictions based on three-variogram functions. This resulted into 120,000 predictions for each variogram function. Like that in Approach-1, in second approach also the authors constructed cumulative frequency distribution followed by relative frequency distribution with six classes having a class interval of 10. The summary for all the three attributes appears in Table 3.

Table 2 Topmost pH prediction accuracy

Prediction accuracy	Frequency count in (%)			Total
	G	S	E	
≥ 90	56.80	27.50	13.20	1950
≥ 80	57.50	28.20	13.25	1979
≥ 70	57.65	28.90	13.35	1998
≥ 60	57.65	28.90	13.35	1998
≥ 50	57.65	28.90	13.35	1998
≥ 40	57.65	28.90	13.35	1998
<40	00.10	0	0	2

drksujatha23@gmail.com

Table 3 Prediction accuracy relative frequency for terrain attributes

Relative frequency out of 120,000 predictions									
PA	pH			EC			OC		
	G	S	E	G	S	E	G	S	E
≥90	64.19	66.96	66.43	17.58	23.41	22.54	20.83	23.70	23.43
≥80	67.94	68.65	68.17	31.35	41.82	41.49	34.27	37.46	37.75
≥70	68.91	68.97	68.49	42.90	55.66	54.44	45.08	48.23	50.30
≥60	68.97	69.25	68.49	60.82	66.59	65.83	61.95	64.84	62.30
≥50	68.97	69.25	68.49	61.66	67.46	65.86	61.75	65.03	62.50
<50	31.03	30.75	31.51	38.34	32.54	34.14	38.25	34.97	37.50

pH predictions showed favorability towards Spherical model in terms of higher frequency for greater than 50% accuracy criteria. The results of EC and OC further justify the existence of higher frequency of Spherical model for accuracy criteria greater than 50%. Thus, authors concluded from this phase of model selection that Spherical variogram model is comparatively better to Exponential and Gaussian. Therefore, the next phase after model identification was to select best NSR combinations using first and second approach with models as Gaussian and Spherical respectively.

3.2 Approaches to Variogram Parameter Selection

The next analysis carried out is to select the best combination of nugget, sill and range that is NSR combination. The authors computed the frequency of a combination contributing prediction accuracy greater than 90% with Gaussian model in first approach. This resulted in higher frequencies for the NSR combinations N1S3R2, N1S3R3, N1S1R3, and N1S1R2. In the second approach, the authors analyzed 120,000 predictions for each of the three-variogram models based on prediction accuracy classes and relative frequencies for each 60 NSR combinations. This resulted into N1S3R2, N1S3R3, N1S2R3, and N1S2R2 as the top four combinations. Since, distance, sample size, and distribution of terrain attributes vary over a period, therefore a single combination all the time may not report nearest prediction as well as its accuracy hence authors selected top four NSR combinations for weight generation.

3.3 Validation and Performance of Model

The authors then validated the model using one third of the total population set consisting of 167 villages. Both Gaussian and Spherical models participated for

drksujatha23@gmail.com

Table 4 Comparative performance of variogram model at validation phase

Model	pH			EC			OC		
	MIN	AVG	MAX	MIN	AVG	MAX	MIN	AVG	MAX
Gaussian	76.89	96.43	99.96	0.0	46.89	99.56	0.0	52.03	99.6
Spherical	83.52	96.87	99.94	2.92	61.11	99.48	1.93	63.00	99.65

validation. The nugget function selected is N1 with two sill functions as S1 and S3 for Gaussian, S2 and S3 for Spherical variogram. Likewise, two common range functions for both the phases are R2 and R3. For each terrain attributes, system performs 8400 experiments. The minimum accuracy of the attributes with Approach-1 is 76.89% for pH and 0% for EC and OC respectively. On the contrary, the minimum prediction accuracy with Approach-2 is 83.52%, 2.92% and 1.93% for the three attributes individually. Table 4 presents the performance summary for the two models. Thus from the validation phase it concluded that there is no substantial difference between the average accuracy of pH for both the approaches, while there is a significant increase in the average frequency of EC. The rise is from 46.89 to 61.11% followed by OC that rises from 52.03 to 63.00%. This indicated that the overall accuracy of predictions with Spherical model is superior.

3.4 Proposed Terrain Attribute Prediction Model (TAPM)

The flow of the TAPM algorithm the authors determined and implemented in R platform along with R Studio.

$$\gamma(h) = \frac{1}{2n}\sum_{i=1}^{n}[z(x_i) - z(x_i + h)]^2 \tag{1}$$

$$\gamma_{2(h)} = \begin{cases} C_0 \left[\frac{3}{2}\frac{h}{a_0} - \frac{1}{2}\left(\frac{h}{a_0}\right)^3\right], & for\, h \le a_0 \\ C_0 & for\, h > a_0 \end{cases} \tag{2}$$

$$\sum_{i=1}^{n} \lambda_i(x_0) z(x_i) = f_0(vi, z) \tag{3}$$

$$z^{\wedge}(x_0) = \sum_{i=1}^{n} \lambda_i(x_0) z(x_i) \tag{4}$$

The system first reads a file-containing list of villages, calculates its count, segregates prediction and predictor villages and computes empirical variogram using Eq. 1. Next, the system calculates Nuggets, Sills, Ranges for four NSR

drksujatha23@gmail.com

alternatives and fits Spherical variogram model Eq. 2. The process then calculates weight and performs predictions using Ordinary Kriging as per Eqs. 3, 4. Final task is calculation of percentage accuracy of four predictions and select maximum prediction.

4 Conclusion and Future Work

Validation approach reveals significant increase in prediction accuracy of Terrain Attribute Prediction Model with second approach accounting to 96.88% for pH, 61.11% for EC and 63.03% for OC. The subjectivity of variogram modelling for this region has overcome using Ordinary Kriging with Spherical Variogram using Nugget as Minimum Variance, Sill as Average and Maximum Variance, Range as Average and Maximum Distance. The model achieves an average accuracy of more than 74%. Thus, authors propose a new dimension for modelling geo-spatial variability of southern Gujarat terrain attributes. The researcher has designed a GIS as a foundation for south Gujarat region to benefit the farmers [17]. They can know the terrain information of their village, predict a land attribute for unvisited locations and visualize maps. Thus, it is a basis for future real-time prediction by linking with GPS device, utilize agricultural prices, meteorological data and there by predict cropping and fertilization schemes, disease mapping, its association and effects of fertilization.

References

1. Zheng S (2010) Crop production on acidic soils: overcoming aluminium toxicity and phosphorus deficiency. Anal Bot 106(1):183–184
2. Thakor K, Dharaiya N, Singh V, Patel A, Mehmood K, Kalubarme M (2014) Soil resources information system for improving productivity using geo-informatics technology. Int J Geosci 5(8):771–794
3. Dey P, Karwariya S, Bhogal N (2017) Spatial variability analysis of soil properties using geospatial technique in Katni district of Madhya Pradesh, India. Int J Plant Soil Sci 17(3):1–13
4. Patel PL, Patel NP, Patel HP, Gharekhan A (2014) Study of basic soil properties in relation with micronutrients of Mandvi tahsil near coastal region of Kutch district. Int J Sci Res 3 (6):28–31
5. Tailor J, Lad K (2017) Assessing geo-spatial distribution of soil profile: a study of Bardoli, Mandvi and Umarpada talukas. Int J Eng Technol Sci Res 4(11):194–199
6. Dasgupta, Make it in India. 7 March 2016. https://www.geospatialworld.net/article/make-it-in-india/. Accessed 22 Oct 2017
7. Bickelhaupt D, Schmedicke R (2017) Soil pH: what it means? College of Environmental Science and Forestry. http://www.esf.edu/pubprog/brochure/soilph/soilph.htm. Accessed 23 Oct 2017
8. Tailor J, Gulati R, Tailor R (2015) Application of ordinary kriging on educational data using FOSS4G: R platform. In: OSGEO-FOSS4G proceedings, Dehradun

drksujatha23@gmail.com

9. Tailor J, Lad K, Gulati R (2017) Measuring spatial correlation of soil pH and Fe using theoretical variograms. Int J Comput Sci Netw 6(5):533–538
10. Saved E, Omran E (2012) Improving the prediction accuracy of soil mapping through geostatistics. Int J Geosci 3(1):574–590
11. Poshtmasari HK, Sarvestani ZT, Kamkar B, Shataei S, Sadeghi S (2012) Comparision of interpolation methods for estimating pH and EC in agricultural fields of Golestan province. Int J Agric Crop Sci 4(4):157–167
12. Burnham K, Anderson D, Huyvaert K (2011) AIC model selection and multimodel inference in behavioral ecology: some background, observations, and comparisons. Behav Ecol Sociobiol 65:23–35
13. Heuvelink GM, Webster R (2001) Modelling soil variation: past, present, and future. Geoderma 100(3–3):269–301
14. Tailor J, Gulati R (2015) Comparing prediction accuracy of OK and RK for the soils of Surat talukas. In: IEEE TIAR-2015, Chennai
15. Rathod R, Kanet G, Sardhara N, Solanki H, Vaghani B, Tailor J (2013) Application of ordinary kriging on educational datasets. Natl J Syst Inf Technol 7(2):91–98
16. Mulatani K, Mohammad A, Jadav M, Tailor J (2016) Analyzing Spatial Autocorrelation in distribution of soil chemical properties using Moran'I and Variogram for talukas of Surat district. Natl J Syst Inf Technol 9(2):97–107
17. Tailor J, Lad K (2017) Terrain mapping for South Gujarat: A GIS based solution for geo-Community. Int J Future Revolut Comput Sci Commun Eng 3(10):230–233

drksujatha23@gmail.com

Sentiment Analysis of Live Tweets After Elections



Palak Baid and Neelam Chaplot

Abstract Election plays an essential role in choosing the leader and deciding the future of the country for next few years. The proliferation of micro-blogging messages or tweets around the elections can be used to predict the sentiments of a person. Using text analysis different opinions and emotions can be identified and that concept is known as Sentiment Analysis or Opinion Mining. As the UP elections were completed and a lot of tweets were available in the research, live tweets were collected for five days during the elections. After collecting the tweets various operations were performed on the tweets and then analysis was done on the tweets to identify the sentiments of the people after election. The tweets were collected specifically related Mr. Yogi Adityanaththh.

Keywords Sentiment analysis · Opinion mining · Tweets · R
YogiAditynath · U.P. election · Machine learning · Artificial intelligence

1 Introduction

Sentiment analysis, also known as Opinion Mining has attracted an augmented captivation. It is an arduous challenge for technologies involving language to achieve valid and logical results. There exist varieties of tasks which classify the data written in a natural language spontaneously into a positive or negative feeling, opinion, emotion or subjectivity which may sometimes be so complicated that even different human annotators are sometimes not willing to agree on the category to be assigned to the text. Personal interpretation by any individual may have discrepancies which may depend upon cultural factors and also upon an individual's experience. Also in short texts if not written properly, the task to interpret that text

P. Baid · N. Chaplot (✉)
Jaipur Engineering College and Research Centre, Jaipur, Rajasthan, India
e-mail: neelam.chaplot@gmail.com

P. Baid
e-mail: palakbaid.95@gmail.com

© Springer Nature Singapore Pte Ltd. 2019

V. S. Rathore et al. (eds.), *Emerging Trends in Expert Applications and Security*,
Advances in Intelligent Systems and Computing 841,
https://doi.org/10.1007/978-981-13-2285-3_36

drksujatha23@gmail.com

also becomes difficult rather strenuous such as in the case of comments or tweets or messages on social networks like Twitter or Facebook.

Twitter is a popular micro-blogging website. We usually come across with multiple tweets on daily basis. Tweets are used to express an emotion. Each tweet is 140 characters in length. With the help of tweets we can recognize many factors like location, content, subject, disaster, political issues, etc.

Analysis of these tweets is a complex and challenging task as the texts is short in the form of a sentence or a news headline rather than a document. The language used in such tweets and to create catchy headlines is very informal with hashtags, emoticons, punctuation, abbreviations, jargons, misspellings, slangs, URLs, and genre-specific terminology which are a way of tagging for Twitter messages.

The main aim of this research is to perform automatic analysis of the tweets to identify the sentiments of the people after the elections. For these purpose live tweets were extracted from Twitter after U.P. election and its analysis was done using R Language.

2 Literature Review

Chin et al. in his paper [1] has analyzed 300,000 tweets. The keywords used were "politics" and "political candidates" or full names of political candidates. They used 2000 tweets to train the classifier and remaining to test the same. They have tested using three algorithms: Support Vector Machine, Nearest Neighbor, and Naive Bayes classification. Overall, Support Vector Machine was most accurate and K-Nearest Neighbor the least and in terms of efficiency Naive Bayes was the best in terms of time.

Murthy in his paper [2] identified what roles do tweets play in political elections. In his paper he concluded that the tweets are more reactive than predictive. He found out that electoral success in not at all related to the success on Twitter and that various social media platforms were used to increase the popularity of a candidate.

Amolik et al. in his paper [3] created the dataset using twitter posts of movie reviews and related tweets about those movies. He used different classifiers like Naïve Bayes, Support vector machine, Ensemble classifier, k-means and Artificial Neural Networks. The results show that 75% accuracy was given by SVM.

Khan et al. in this paper [4] they aimed to improve the efficiency of sentiment classification by proposing a new algorithm which performs several data transformations and uses a combination of different text classification algorithms. The proposed classification algorithm classifies the tweets on the basis of improved polarity classifier, emoticon classifier, and SentiWordNet classifier. They pointed out that Ye and Wux [5] in their paper could not explain the basis as to why the tweets became so popular and how they believed a tweet to be generated by a virus or a worm. In the research by Fu et al. [6] the Chinese language dataset was used for testing and and the specifics about how the manual audit of an online retrospection

drksujatha23@gmail.com

was considered to be biased was also missing in this paper. The research by Bifet et al. [7], was also found to be lacking as the authors used only two classes—positive and negative for sentiment categorization. Their results show that they proposed framework achieved an accuracy of 85.7% which was an improvement to previous similar works.

Thelwall et al. in his paper [8] worked on the question that whether the surges of reactions expressed on twitter platform to express sentiments are associated with the increase in the strength of emotions expressed. He concluded that there is strong evidence proving that negative sentiments are more central but for positive sentiment there is some evidence that the same is true.

Tumasjan et al. in his paper [9] has analyzed 104,003 tweets which referenced to either a politician or a political party. They have analyzed the addressivity and retweets as a pointer for the swapping of ideas. The retweets also contain valuable information such as link to other websites. They have also analyzed the rate of retweets which implies that more the rate of retweeting the tweets, more valuable the information. In the same way more the number of tweets represent the voter's preference. So their final conclusion was that the Twitter can be referred and considered as a valid suggestion for a political opinion.

3 Technology and Experiment

3.1 R Language

R provides software environment for graphical and tabular representation, statistical calculations, and reporting and it is the open source programming language. This language was created at the University of Auckland, New Zealand. It is the language which allows programs to have branching and looping along with functional programming using multifarious functions. R is freely available and different versions are provided with various operating systems like Mac, Linux, and Windows. Ri386 3.3.3 is the version used in our experiment. It is a simple and effective programming language. It can effectively handle and store data. R provides various analytical and integrated collection of tools for text analysis and also provides facilities for graphical and tabular representation of data.

3.2 Libraries Used

ROAuth. It provides an interface to the OAuth 1.0 specification and also allow users to authenticate with the help of OAuth to the server preferred by the user.

twitter. It provides an interface to the Twitter web API.

drksujatha23@gmail.com

httr. It is a useful tool for working with HTTP protocol ordered by HTTP verbs.

plyr. It is a set of tools that disintegrates a significant problem down into manageable pieces, perform different functions on each piece and then put all the divided pieces back together. This package helps to solves a common set of problems.

stringr. It is the most simple and facile set of wrappers around the "stringi" package. All the functions and argument names are compatible and consistent, all functions used in this package deal with "NA"s and zero length vectors and the output of one function can be given as the input to another.

plotrix. It is used for variety of plots, distinct and unique labelings, axes and color scaling functions.

tm. It provides a framework for all those applications which are working on opinion mining within R.

wordcloud. It is used to print pretty word clouds.

tidyverse. It is a set of packages that share common data representations and "API" design. This package is designed so that it becomes effortless to install and load multiple "tidyverse" packages in a single step.

lubridate. It has a consistent and facile syntax that makes operations and the usage of dates easy.

tidytext. It is used in text mining for sentiment analysis and word processing using "dplyr", "ggplot2", and other tidy tools.

broom. This is used to convert statistical data into data frames which can then be easily combined, reshaped and then finally processed by tools such as "dplyr", "tidyr" and "ggplot2".

scales. It helps to map data to graphical scales and provide methods for automatically determining breaks and labels for axes and legends.

3.3 Experimental Procedure

Experimental procedure described in Fig. 1 was performed for results. Details of various task performed during the experiments are described in details.

Twitter Authentication and R Installation. In order to extract tweets Twitter application is required and hence a twitter account is opened. After login new application is created on the Twitter platform for further experiment.

After installing R, various packages required are installed and loaded. R script is written to authenticate the twitter account. Authentication creates a link between the programming language and the Twitter application.

Extracting Tweets. Tweets are extracted using the R script. Hashtags are used to extract the tweets. The function searchTwitter() was used to extract the tweets after UP election. The tweets were extracted by providing hashtag #Yogi_Aditynath.

drksujatha23@gmail.com

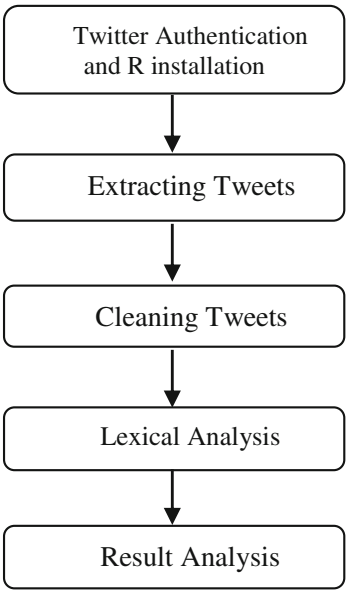


Fig. 1 Experimental procedure for analyzing tweets

The dates of the tweets extracted were from March 24, 2017 to March 28, 2017. The tweets extracted were only of English language. One thousand five hundred tweets were extracted and saved.

Cleaning Tweets. After tweets are extracted, they are cleaned, noise, emoticons, URLs, extra spaces, hashtags, are removed from the tweets. After cleaning the tweets, are converted to data frames for further use.

Lexical Analysis. Analyses of sentiments of words from the extracted tweets were done. Datasets of both positive and negative words were used to identify the positive and negative emotions and sentiments of those tweets.

Result Analysis. A Pie chart, Word cloud, and histogram are created for the analysis purpose. The pie chart created represents the total positive and total negative tweets in our experiment. Word cloud is basically the cloud formed by words changing. The decrease in font size is due to decrease in the intensity of the words in the extracted tweets.

The Histogram of the score of the tweets seen in Fig. 2 shows that there are more negative tweets then positive tweets for the hashtag #Yogi_Aditynath that were extracted.

It is observed that 44.44% tweets were positive and 55.56% tweets were negative.

The pie chart in Fig. 3 shows the positive and negative sentiments of the tweets.

drksujatha23@gmail.com

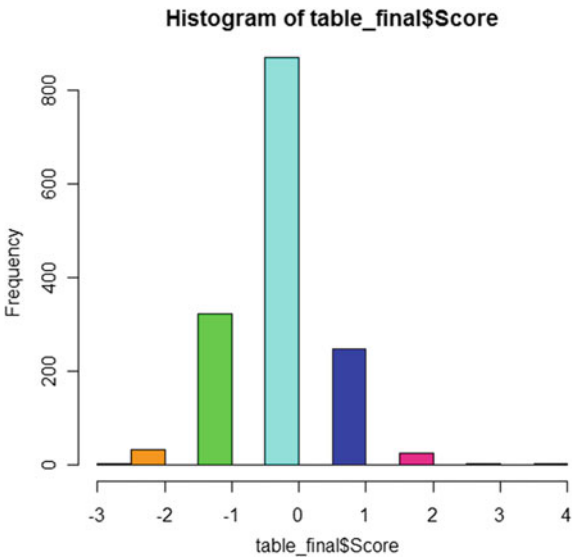


Fig. 2 Histogram of final score of all tweets

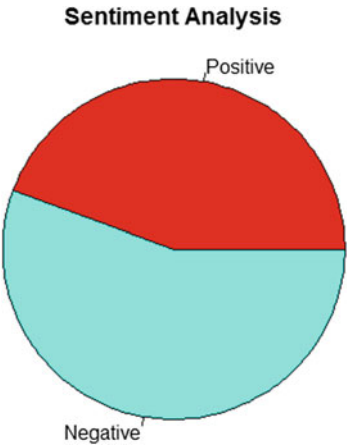


Fig. 3 Pie chart depicting the overall positive and negative scores

The words frequently used in the tweets are shown in the word cloud in Fig. 4. It can be seen with the help of size of the words in the word cloud it can be identified that which words very used more frequently than others.

drksujatha23@gmail.com



Fig. 4 Word cloud depicting the mostly used keywords in Tweets

4 Conclusion and Future Scope

It is observed that overall sentiments expressed in the tweets towards the newly elected Chief Minister of Uttar Pradesh "AdityanathYogi" in March as depicted in the pie chart and bar graph is more negative than positive. Around 44.44% of the tweets were positive and 54.56% of the tweets are negative. The word cloud generated shows that Yogiadityanath is the most used words in the tweets followed by station and police as the next most tweeted words.

The results generated by the analysis are different from our expectations and further studies are required to be done to identify the relationship and sentiments of the persons tweeting about the elections.

References

1. Chin D, Zappone A, Zhao J (2016) Analyzing Twitter sentiment of the 2016 presidential candidates
2. Murthy D (2015) Twitter and elections: are tweets, predictive, reactive, or a form of e buzz? Inf Commun Soc 18(7):816–831
3. Amolik A, Jivane N, Bhandari M, Venkatesan M (2016) Twitter sentiment analysis of movie reviews using machine learning techniques, School of Computer Science and Engineering, VIT University, Vellore, vol 7(6), pp 2038–2044
4. Khan FH, Bashir S, Qamar U (2014) TOM: Twitter opinion mining framework using hybrid classification scheme, Computer Engineering Department, College of Electrical and Mechanical Engineering, National University of Sciences and Technology (NUST), Islamabad, Pakistan, vol 57, pp 245–257

drksujatha23@gmail.com

5. Ye S, Wu SF (2010) Measuring message propagation and social influence on Twitter.com. In: Bolc L, Makowski M, Wierzbicki A (eds) SocInfo 2010, LNCS 6430. Springer-Verlag, Berlin, Heidelberg, pp 216–231
6. Fu X, Guo Y, Guo W, Wang Z et al (2012) Aspect and sentiment extraction based on information-theoretic co-clustering. In: Wang J, Yen GG, Polycarpou MM (eds) ISNN 2012, Part II, LNCS 7368. Springer-Verlag, Berlin, Heidelberg, pp 326–335
7. Bifet A, Holmes G, Pfahringer B (2011) MOA-TweetReader: real-time analysis in twitter streaming data. In: Elomaa T, Hollm'en J, Mannila H (eds) DS 2011, LNCS 6926. Springer-Verlag, Berlin, Heidelberg, pp 46–60
8. Thelwall M, Buckley K, Paltoglou G (2011) Sentiment in Twitter events, statistical cybermetrics research group. J Am Soc Inform Sci Technol 62(2):406–418
9. Tumasjan A, Sprenger TO, Sandner PG, Welpe IM (2010) Predicting elections with Twitter: what 140 characters reveal about political sentiment. In: Fourth international AAAI conference on weblogs and social media, pp 178–185

drksujatha23@gmail.com

Smart Innovation Regarding Bringing Kitchen Food Items in the Kitchen by Automatically Informing the Shopkeeper by Using GSM 900 Board and Arduino Uno R3 Board with Proper Programming



Vijay Kumar, Vipul Sharma and Avinash Sharma

Abstract Smart innovation can be done from GSM 900 board, Arduino Uno R3 board, and sensors for bringing kitchen food items, i.e., supply of rice, pulses, flour, etc. from shopkeeper by sending message to the shopkeeper automatically. The sensor senses the voltage by weight of the material and if weight is less, low voltage is generated while if weight is high, high voltage is generated by the presence of materials in the vessel. This voltage is given to pin 3 of Arduino Uno R3 board which by proper programming is used to send the SMS (message) to the shopkeeper automatically. Thus, the flour which is if about to empty in my vessel is informed to the shopkeeper automatically and the shopkeeper will send the flour to house.

Keywords Sensors · GSM 900 board · Arduino Uno R3 board
SMS service

Vijay Kumar—Masterminded and created the first stable version of this document.

Vipul Sharma—Created the first draft of this document.

Avinash Sharma—Created the first draft of this document.

V. Kumar (✉)
Aryabhatta Centre for Nanoscience and Nanotechnology,
Aryabhatta Knowledge University, CNLU Campus, Patna, India
e-mail: vijay898627@gmail.com

V. Sharma
Gurukul Kangri University, Haridwar, India
e-mail: vipul.s@rediffmail.com

A. Sharma
MMEC, Mullana, Haryana, India
e-mail: sh_avinash@yahoo.com

© Springer Nature Singapore Pte Ltd. 2019

V. S. Rathore et al. (eds.), *Emerging Trends in Expert Applications and Security*,
Advances in Intelligent Systems and Computing 841,
https://doi.org/10.1007/978-981-13-2285-3_37

drksujatha23@gmail.com

1 Introduction

Nowadays, people are doing research on Internet of things, viz., IoTs. IoTs can be used in many sophisticated tasks such as closing or opening of door of house from remote place, control the power of home appliances from outside home, etc [1]. This work is about bringing kitchen food items like rice, pulses, flour, etc. automatically from shopkeeper when the vessel is about to be emptied. This can be done by proper programming of Arduino Uno R3 board and GSM 900 board with attached sensors. Sensors can be used for change in voltage with respect to change in the weight of the vessel. I have used different voltages directly with the Arduino Uno R3 board.

In the literature survey for this topic, we have found that previous to this paper no one has used analog pin 3 of Arduino Uno R3, for sending of message by sensing of voltage on this pin. This is the main crux of this research topic.

2 Procedure to Do the Task

- We have taken a GSM SIM card and inserted it to the GSM 900 board [2].
- GSM Tx has been connected to pin 9 (serial port) of Arduino Uno R3 board [2].
- GSM Rx has been connected to pin 10 (serial port) of Arduino Uno R3 board [2].
- GSM ground has been connected to the ground of Arduino Uno R3 board [2].
- We are using the analog port 3 of Arduino Uno board for the voltage variation measurements [3].
- We have put the condition for the voltage to send the message, corresponding integer, to be greater or equal than 777 [3].
- GSM 900 board and Arduino Uno R3 board is powered on.
- Arduino software platform is used for programming the Arduino Uno R3 kit to perform the task.
- After writing the software on Arduino Uno R3 kit, we have directly connected pin 3 of analog port to 5 V for showing that the vessel has got emptied.
- After few seconds, Arduino kit senses the voltage and sends message to the number fed.
- Block diagram of the system is given in Fig. 1.

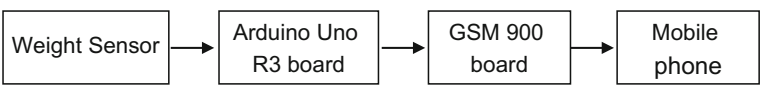


Fig. 1 Block diagram of the system

drksujatha23@gmail.com

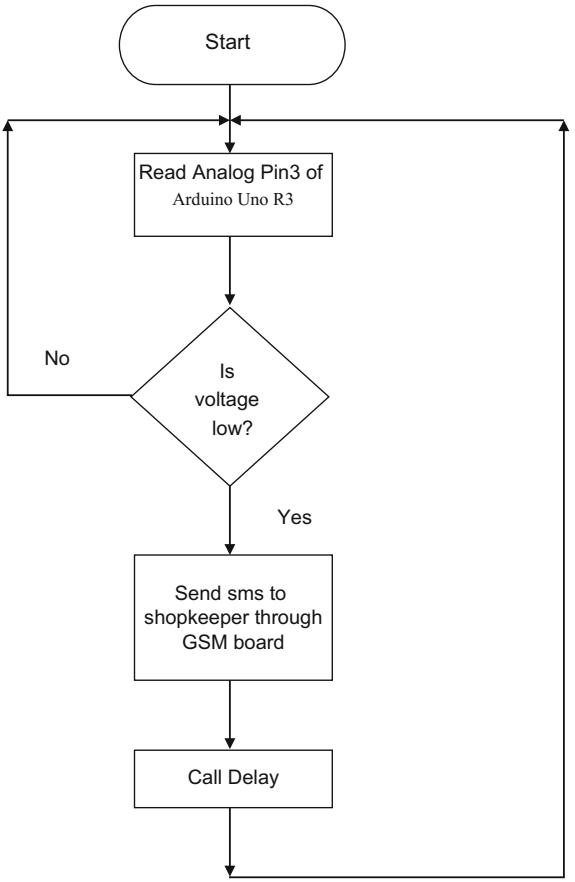


Fig. 2 Flowchart of the algorithm

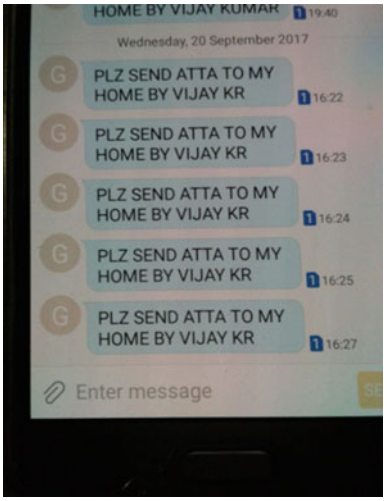


Fig. 3 Photograph showing message on mobile phone

drksujatha23@gmail.com

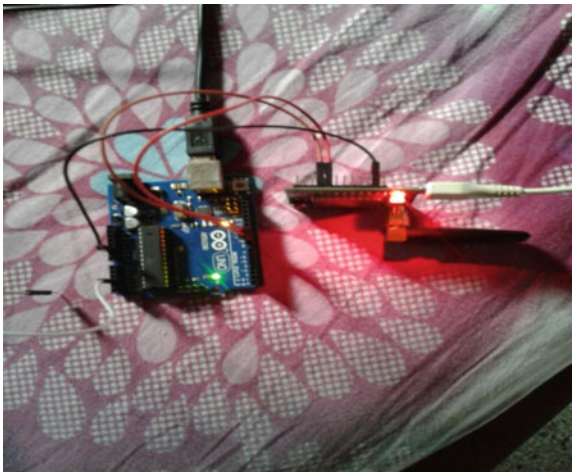


Fig. 4 Photograph of hardware

- Figure 2 shows flowchart of the program algorithm.
- Figure 3 shows snapshot of the message received on the intended mobile number.
- Figure 4 presents photograph of hardware of the system.

3 Conclusion

Our hardware is a success for informing shopkeeper regarding the sending of kitchen food items to one's house. In our above paper, we have used GSM 900 Tx Rx board and Uno R3 board for informing the shopkeeper through SMS about less amount of items in kitchen and sending the items to house. Future work is regarding using sensors for sensing voltage and the other important thing—payment.

References

1. Kosmatos EA, Tselikas ND, Boucouvalas AC (2011) Integrating RFIDs and smart objects into a unified internet of things architecture. Adv Internet Things Sci Res 1:5–12. https://doi.org/10.4236/ait.2011.11002
2. www.circuitstoday.com/interface-gsm-module-with-arduino
3. https://www.arduino.cc/en/Reference/HomePage
4. Lianos M, Douglas M (2000) Dangerization and the end of deviance: the institutional environment. Br J Criminol 40:261–278. https://doi.org/10.1093/bjc/40.2.261
5. Ferguson T (2002) Have your objects call my object. Harvard Business Review, June, 1–7
6. Nunberg G (2012) The Advent of the internet: 12th April, Courses
7. Aggarwal R, Lal Das M (2012) RFID security in the context of "Internet of Things". In: First international conference on security of internet of things, Kerala, 17–19 August 2012, pp 51–56. http://dx.doi.org/10.1145/2490428.2490435

drksujatha23@gmail.com

Sentence Tokenization Using Statistical Unsupervised Machine Learning and Rule-Based Approach for Running Text in Gujarati Language



Chetana Tailor and Bankim Patel

Abstract Sentence tokenization is the foundational step in natural language processing to analyze the sentence. Apart from others, main causes which make the sentence tokenization difficult are quotation marks and the multipurpose usage of punctuation marks especially dot ".". In this paper, a framework has proposed for sentence tokenization for running text in Gujarati language using statistical unsupervised machine learning approach and rule-based approach.

1 Sentence Tokenization and Issues

A sentence is a complete set of words that conveys meaning and on which syntactic rules can be applied [1]. As defined in [2, 3], the role of sentence tokenizer is to identify and segment the sentences made up of words and punctuation marks incorporating the syntactic rule from the text. Sentence tokenization is important for Natural Language Processing applications like creating bilingual Parallel corpora [4], Anaphora resolution [5], POS tagging [6], Machine translation [7], etc. which requires sentence as an input. Therefore, accuracy of sentence tokenization is important as it directly affects the performance of NLP applications [8]. Different punctuation marks listed in [9] are used to identify sentence boundary detection in natural languages. Among these, usually punctuation marks dot, exclamation mark, and question mark are used as sentence end maker [10]: out of these, dot is the most ambiguous as it is used for different purposes like sentence end marker, name initial, decimal point in number, e-mail address, website name, truncated word, etc. [2, 11]. These punctuation marks are used to indicate sentence end markers but in direct speech text, these punctuation marks can occur internally in sentence, and

C. Tailor (✉) · B. Patel
Shrimad Rajachandra Institute of Management and Computer Application,
Uka Tarsadia University, Bardoli, Surat, India
e-mail: chetana.tailor@gmail.com

B. Patel
e-mail: bankim_patel@srimca.edu.in

© Springer Nature Singapore Pte Ltd. 2019

V. S. Rathore et al. (eds.), *Emerging Trends in Expert Applications and Security*,
Advances in Intelligent Systems and Computing 841,
https://doi.org/10.1007/978-981-13-2285-3_38

drksujatha23@gmail.com

also inside the token in case of complex token [12, 13]. Due to the difference in representation of Unicode quotes (“”, ‘’) and ASCII quotes ("", '), the difficulty in sentence tokenization gets increased [8]. Another issue is the identification of closing single quotation mark and apostrophe [12]. Sentence having parentheses and punctuation marks inside the parentheses need to be handled carefully as punctuation marks inside the parentheses are not the actual sentence end marker [14]. Thus, identification of the role of punctuation mark is important for sentence boundary detection and tokenization [2, 8, 11–15]. Different approaches like rule-based approach [16–21], supervised machine learning approach [22–26], and unsupervised machine learning approach [27] have been used by researchers for sentence tokenization.

Authors in [16–18] have used rule-based approach for identifying the role of dot for English language. Bayer [16] has achieved accuracy 98.4% on The Call of the Wild corpus and 93.95% accuracy on The Wall Street Journal corpus. In The Call of the Wild corpus, out of 1640 periods only 1613 periods while in The Wall Street Journal corpus out of 16466 periods, 15470 periods are correctly identified. Left context and right context of the dot are used as features. The difference, as compared to Bayer's approach from Alembic information extraction system [17], is dictionary containing list of abbreviations that is used along with rules to identify the role of dot "." in number, date, time, and abbreviations. Achieved accuracy is 99.1% on Wall Street Journal Corpus. Mikheev [18] has adopted the rule-based approach based on the word to the left or right that is abbreviation or proper noun which are made based on heuristics on unlabelled training data and reported 0.45% error rate on The Wall Street Journal corpus and 0.28% on the Brown Corpus. Merging of this approach with a supervised POS tag [28] reduces the error rates to 0.31% and 0.20%, respectively.

Authors in [19, 20] have adopted rule-based approach for Kannada Language. Parakh et al. [19] have given more attention to disambiguate dot as sentence end marker, abbreviation, and salutation, while Deepmala and Ramakanth [20] have included question mark and exclamation mark along with dot. In [19], authors have developed two lists and threshold derived from corpus that narrows down its scope. List L1 contains valid sentence ending words with a length below a threshold and L2 contains an ambiguous word with the length below threshold. They have achieved 91.33% accuracy and if the rules apply on the same corpus, on which it has been developed, then 99.14% accuracy is achieved. Abbreviations and Verbs lists are used in [20]. They have tested on EMILLE corpus with size 227 KB and achieved 99.2% accuracy. Wanjari et al. [21] have used rule-based approach for Marathi language. Their main focus is to identify the abbreviation using prefix and suffix of the punctuation mark and compare the POS tag of the context for sentence boundary detection.

Riley [22] has used regression tree to classify the dot (.) as sentence end marker or used in abbreviation with contextual features like word preceding and following the punctuation mark ("."), length and case of the word preceding and following punctuation mark dot ("."), punctuation mark after dot ("."), and list of abbreviation with dot ("."). Achieved accuracy is 99.8% on Brown corpus.

drksujatha23@gmail.com

Reynar and Ratnaparkhi [23] and Negi et al. [25] have used a supervised maximum entropy model with consideration of three punctuation marks: dot ".", exclamation marks, and question mark. Authors in [23] have used contextual features like preceding and following word of word containing punctuation mark and preceding and following word's feature like whether it is honorific word or abbreviation. In both the approaches, dictionary having list of abbreviation and honorific words created during the training is used. The differences between both the approaches are that text preprocessing and corpus of [25] are not the same as used in [23]. This approach [23] has achieved 98% accuracy on The Wall Street Journal corpus. Performance of this approach is less as compared to Riley's performance because Riley uses 30% more data as compared to Reynar and Ratnaparkhi's training corpus. However, corpus detail in [25] is not published but achieved accuracy is 99%. Sentence boundary detection system by Palmer and Hearst [24] is based on part of speech probability of three words preceding and following the punctuation marks along with neural network and decision tree classifiers. Achieved accuracy is 98–99% on the Wall Street Journal Corpus. Error occurs due to abbreviations, quotation marks, parenthetical sentences, and ellipsis.

Wong et al. [26] have discussed the ambiguous role of dot, semicolon, colon, comma, exclamation mark, and question mark. The iSentenizer-μ system has been developed by implementing an incremental tree learning architecture to detect the sentence boundary of for Danish, German, English, Spanish, Dutch, French, Italian, Portuguese, Greek, Finnish, and Swedish languages along with the contextual information the current word, preceding word, and following word as features. The system has achieved the 99.8%, 99.81%, and 99.78% accuracy on the Wall Street Journal, Brown, and Tycho Brahe corpus, respectively. Rudrapal et al. [29] have addressed the issues regarding dot, question marks, exclamation marks, and consecutive occurrence of punctuation marks to detect the sentence boundary on social media. They have used two different approaches: rule-based and supervised machine learning approaches. Naive Bayes, conditional random fields, and sequential minimal optimization approaches have been tested with the features such as two preceding words and one following word to the current word. Among them, Naive Bayes machine learning gives 99.6% accuracy that is equivalent to splitta: SVM (https://code.google.com/p/splitta/).

In [27], Kiss and Struck have focused to solve the problem regarding role of dot in a sentence. They have introduced unsupervised machine learning. It uses collocation method to identify the abbreviation, initial, and ordinal numbers. It works on log-likelihood-based classification technique. Average accuracy on 11 different languages is 98.74% on newspaper corpora.

From the above literature survey, authors have observed that the followings are the key issues that must be addressed while developing the sentence tokenizer for natural language: First is **disambiguate the role of dot "."**: Dot "." is used as sentence end marker, in abbreviation either as initial of names or truncated words, decimal point in number, web URL, and e-mail address. Second is a **parenthetical expression**: A parenthetical expression is word or word phrase added within a sentence, which does not change the actual meaning and structure of the sentence

drksujatha23@gmail.com

but qualify the meaning of the sentence. Sentence should not be segmented based on parenthetical expression. Third is **quotation marks**: The quotation marks contain expression like number with decimal point, abbreviation, etc. with punctuation marks should not be tokenized. Fourth is r**epetition of punctuation marks**: Identification of sentence boundary when sentence containing ellipse, more than one occurrence of exclamation marks or question marks, or any combination of punctuation marks.

Authors are motivated for sentence tokenization for Gujarati language because it is the foundational step in NLP applications [4–8] that affects the result of such applications. From the above literature survey, authors have observed that rule-based approach has dominancy [16–21] for English as well as Kannada language but rules are derived from corpus which increases the corpus dependency, however, unsupervised statistical approach: Punkt [27] has been tested on 11 different languages and mean accuracy on 11 different languages is 98.74% on newspaper corpora. This model is also used for Gujarati language in [3] using NLTK; however, authors have neither published the accuracy nor about the corpus detail. Due to the lack of available sentence tokenizer for Gujarati language with good accuracy, authors are motivated for further study and for the same they have combined the rule-based and unsupervised statistical approach.

2 Sentence Tokenizer for Running Text in Gujarati Language

Nearly, 50 million people living in western part of India speak Gujarati language and 65.5 million people in all over the world are Gujarati speakers that make Gujarati language the 26th most spoken native language in the world. General structure for sentence tokenization is shown in Fig. 1. Corpus for running text has been prepared incorporating common issues observed in literature survey by collecting online Gujarati news articles that has 2.63 MB of size. Training corpus contains four different news article categories, namely, Crime, Sports, Business, and Politics along with diversity in the sentence with different punctuation marks in news articles. General framework of this system is given in figure number 1. It is

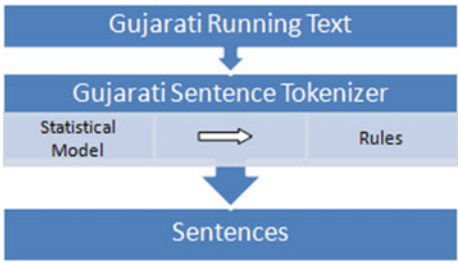


Fig. 1 Sentence tokenization framework

drksujatha23@gmail.com

implemented using Python language. Punkt model for Gujarati language has been trained on 2.63 MB of training corpus using NLTK [30]. Model handles e-mail address, number with decimal point, question marks, and exclamation marks. Authors have studied and analyzed the Punkt model for Gujarati language, which fails to address the following issues for Gujarati running text: First issue is an abbreviation made up of initials with space in between the initial or truncated word within a sentence. Second is sentence with parenthetical expression. Third is more than one sentence within quotation marks, and forth issue is a sentence having two words with dot as punctuation marks are not identified. Example: રામ ભજો. રામ બધાનું સારુ કરશે. [rām bʱəʤo. rām bədʱāɳuɳ sāru kərəʃe.] [Worship the Raama. Raama will make everything good for all.] In example, sentence રામ ભજો. [rām bʱəʤo.] [Worship the Raama.] has only two words which is not identified by the model. Instead of giving two sentences as tokens, it gives only one sentence as a token. Fifth issue is the identification of sentence boundary in the sentence having ellipse, ordered, and unordered list.

From the above listed five issues, four issues are common to general issues found in natural languages. Here, authors have focused to solve the issues: initial in abbreviations, quotation marks, parenthetical expression, and ordered list for that authors have developed and implemented rules. After text processed by model, developed rules are applied in the following order: First is **identification of the abbreviation composed by name initials**: Users use English and Gujarati language alphabets including English character transliterated into Gujarati. To identify the initial of the name, two rules are defined: one rule to handle the initial of the name made up of English alphabats and another for Gujarati and transliterated English alphabats into Gujarati. These rules handle abbreviations containing initials as discussed in above-observed issue number 1. Second is a **parenthetical expression**: In running text, to provide the meaning clarity, parenthetical expressions are used. Observed issue number 2 in text having parenthetical expression that has punctuation marks is resolved. Third is to **knob the order list**: Order list containing numbers (where number can be either in Gujarati language or in English language) or alphabets with different formats—closing parenthesis, opening and closing parentheses, closing rectangle bracket, opening and closing rectangle brackets, and order list without parentheses and rectangle brackets. This rule handles the above-observed issue number 5. Fourth is **Quotation marks ambiguity resolution**: Quotation marks are used to provide the importance to the word in the sentence and to note the direct speech. However, usage of quotation marks as well as representation of quotation marks “”, ‘’, "", and " creates ambiguity. This rule handles quotation marks with multiple sentences, quotation marks at the end of the sentence, within a sentence, nested quotation marks in the sentence as well as variation in quotation mark symbol representation as listed in observed issue number 3. Exceptions in the first rule for the collocation pattern like “હા. [ɦā.] [Yes.], ના. [nā.] [No.]” are also handled as they are mostly used as sentence end marker not as an initial.

drksujatha23@gmail.com

Table 1 Result of sentence tokenizer for the Gujarati language

Sr. no.	Domain	Total sentences	Error(s)	Accuracy (%)
1.	Business	354	1	99.72
2.	Crime	296	8	97.3
3.	National category of EMILLE corpus	288	3	98.96
4.	Politics	394	1	99.76
5.	Sports	383	0	100
6.	Technical	346	0	100
7.	Vaividhya	526	2	99.62
Total		2587	15	99.34

3 Experimental Analysis

Authors have developed their own corpus for testing the developed Gujarati sentence tokenizer system. Developed corpus for testing has 140 news articles containing 20 articles from each of categories Business, Crime, Political, Sports, Technical, National category of EMILLE corpus, and Vaividhya. In this testing corpus, totally 3460 punctuation marks are present excluding comma, semicolon, and colon. Out of 3460 punctuation marks, totally 3262 dots, 28 question marks, 6 exclamation marks, 91 quotation marks, and 71 brackets are in the testing corpus. Among 3262 dots, there are 342 dots used in abbreviations containing name initials, 27 in truncated words, 188 decimal numbers, 40 in repetition of punctuation marks, 2 dots in e-mail address, 3 in address, and rest of the dots are used in quotation marks, and sentence end markers. 30 order lists are included in corpus. Authors have identified 15 errors out of 2587 sentences. Mean accuracy achieved on this seven domain of the corpus is 99.34%. Errors are found in this Gujarati sentence tokenization due to ellipse and truncated words. Accuracy of this system on different categorical articles is shown in Table 1.

4 Conclusion

Applying Punkt unsupervised learning followed by rules with precise features to the sentence tokenization task is very successful for running text in Gujarati Language where Dot ".", exclamation mark "!", and question marks "?" are considered as sentence end markers. This combined approach is very effective as there is no use of POS tag as well as no extra resources like abbreviations list or list of verbs are used. This approach is quite robust as tested on 140 different articles. By adding

drksujatha23@gmail.com

specific features targeted to repetition of punctuation marks and truncated word, we believe that we could obtain even better results. Authors are working to create truncated words list.

References

1. Andersen S (2016) Sentence types and functions. (2014) Internet: http://www.sjsu.edu/writingcenter/handouts/Sentence%20Types%20and%20Functions.pdf. Accessed 22 Dec 2016
2. Grefenstette G, Tapanainen P (1994) What is a word, what is a sentence? problems of tokenization. In: 3rd international conference on computational lexicography, pp 1–11
3. Hardeniya N et al (2016) Tokenizing text and wordnet basics. In NLP: Python & NLTK, Packt
4. Zhu L, Wong DF, Chao LS (2014) Unsupervised chunking based on graph propagation from bilingual corpus. Sci World J 2014, Article ID 401943:1–10
5. Zheng J et al (2012) A system for coreference resolution for the clinical narrative 19(4)
6. Manning C (2011) Part-of-speech tagging from 97% to 100%: is it time for some linguistics? Comput Linguist Intell Text Process 6608:171–189
7. Walker D et al (2001) Sentence boundary detection: a comparison of paradigms for improving MT quality. In: Proceedings of the MT Summit VIII, Santiago de Compostela, Spain
8. Read J et al (2012) Sentence boundary detection: a long solved problem? In: Proceedings of COLING 2012: posters, pp 985–994, Mumbai, December 2012
9. Straus J (2008) Punctuation. In: The Blue Book of Grammar and Punctuation, Chapter 3, pp 052–068
10. Manning C, Schütze H (2002) Corpus-based work. In: Foundations of statistical natural language processing. The MIT Press, London, England, Chapter 4, Section 4.2.4, pp 134–136
11. Chithra C, Ramaraj E (2016) Heuristic sentence boundary detection and classification. Int J Emerg Technol 7(2):199–206
12. Indurkhya N, Damerau F (2010) Sentence segmentation. In: Handbook of natural language processing. Chapman & Hall/CRC, Taylor & Francis Group, Chapter 2, Section 2.4.1, pp 023–024
13. Maurya S et al (2016) Gender and number identification for Gujarati word: rule-based approach. NJSIT 9(2):1–7
14. Dias K (2015, August 1) Pragmatic segmenter [Online]. https://www.tm-town.com/natural-language-processing. Accessed 23 Dec 2016
15. Nunberg G (1990) The linguistics of punctuation. Stanford, CA: C.S.L.I. Lecture Notes, vol 18
16. Bayer S et al (1998) Theoretical and computational linguistics: toward a mutual understanding. In: Using computers in linguistics: a practical guide, 238–253
17. Aberdeen J et al (1995) MITRE: description of the Alembic system used for MUC-6. In: The Proceedings of the 6th MUC, Columbia, Maryland, November 1995, pp 144–155
18. Mikheev A (2000) Tagging sentence boundaries. In: Proceedings of NAACL, pp 264–271, May 2000
19. Parakh M et al (2011) Sentence boundary disambiguation in Kannada texts. In: Special volume: problems of parsing in Indian languages, pp 17– 19, May 2011
20. Deepmala N, Kumar R (2012) Sentence boundary detection in Kannada language. Int J Comput Appl (0975–8887), 39(9)
21. Wanjari N et al (2015) Sentence boundary detection for Marathi language. Procedia Comput Sci 78:550–555 Elsevier, Science Direct

drksujatha23@gmail.com

22. Riley M (1989) Some applications of tree-based modeling to speech and language indexing. In: Proceedings of the DARPA speech and natural language workshop, Pennsylvania, February 1989, pp 339–352
23. Reynar J, Ratnaparkhi A (1997) A maximum entropy approach to identifying sentence boundaries. In: Proceedings of the 5th conference on applied natural language processing, pp 16–19
24. Palmer D, Hearst M (1997) Adaptive multilingual sentence boundary disambiguation. Comput Linguist 23(2):242–267
25. Negi P et al (2010) Sentence boundary disambiguation: a user friendly approach. Int J Compute Appl 7(8):033–037
26. Wong D et al (2014) iSentenizer-μ: multilingual sentence boundary detection model. Sci World J © 2014 Wong DF et al. http://dx.doi.org/10.1155/2014/196574
27. Kiss T, Strunk T (2006) Unsupervised multilingual sentence boundary detection. Comput Linguist 32(4):485–525
28. Mikheev A (2002) Periods, capitalized words, etc. Comput Linguist 28(3):289–318
29. Rudrapal D et al (2015) Sentence boundary detection for social media text. In: ICON-2015, pp 91–97
30. Willy et al (2016) Natural language toolkit: Punkt sentence tokenizer [Online]. http://www.nltk.org/_modules/nltk/tokenize/punkt.html. Accessed Nov 2016

drksujatha23@gmail.com

A Hybrid Approach to Authentication of Signature Using DTSVM



Upasna Jindal and Surjeet Dalal

Abstract Signature verification is a widely developed area of research for authentication. A biometric method is used for identification and verification. A unique characteristic of a human like palm, iris, voice, fingerprints, etc. are being used for authentication. Generally, in examination, banking, and any other transcation, two types of signatures are used: (a) handwritten and (b) using any digital device like stylus. Signature verification is the most accepted technique to overcome the problem of forgery from the signature. The main aim of our paper is to provide the authentication of signature using support vector machine technique. As we all know that SVM has many different kinds of function used as kernel such as linear kernel, radial basis kernel, Gaussian kernel, etc. For all these kernels, various types of parameter selection algorithm are available. In this research, we propose a hybrid algorithm, i.e., decision tree support vector machine (DTSVM) for multiple class classification. Using the decision tree algorithm, our DTSVM effectively overcomes the forgery factory from the signature with respect to other effective techniques. The previous experimental results explain that the proposed algorithm is able to find the significant results for skilled forgery in terms of false acceptance ratio, false rejection ratio, and equal error ratio and has better classification accuracy as compared with other algorithms applied.

1 Introduction

In the last few decades, plenty of ideas have been proposed, developed, and tested in the area of signature verification. There are two main approaches for handwritten signature: offline in which signature is done using pen and paper, generate static image, and another approach is online which generates dynamic image of the

U. Jindal (✉) · S. Dalal
SRM University, Sonipat, Haryana, India
e-mail: upasnajindal@gmail.com

S. Dalal
e-mail: profsurjeetdalal@gmail.com

© Springer Nature Singapore Pte Ltd. 2019

V. S. Rathore et al. (eds.), *Emerging Trends in Expert Applications and Security*,
Advances in Intelligent Systems and Computing 841,
https://doi.org/10.1007/978-981-13-2285-3_39

drksujatha23@gmail.com

signature, which is captured through any electronic device. Nowadays, online signature is considered as the secured as compared to the offline. For signature, a biometric approach is highly flourishing in the society and having much importance in terms of security. If we want signature security, then the verification is required for both types of signature. Significant amount of work has been already done on online and offline signatures. Lots of researcher have produced more than 90% accuracy in online signature verification; here, we proposed an integrated approach where offline and online signatures generate equal accuracy [1–4].

Figure 1 depicts the integrated signature verification system. There are three different phases in the block diagram. The first phase is preprocessing for removal of unwanted factors from the signature before produce to the next phase. Second phase is the feature extraction where local, global, and transit features are taken out from signature image and decision tree SVM algorithm has been used to verify an image. The aim of our work is to present an integrated verification system to improve the quality of signature, by matching scores of individual using a hybrid technique, i.e., decision tree support vectors machine [5–8]. To cut down the error rate from signature, our proposed technique is the good solution for skilled forgery detection. The overall performance of the system depends on the scores generated by signature. The accuracy is degraded and performance gained due to equality shortage of the individual classifier [9–14]. To overcome the forged factor from signature, we present a hybrid technique, i.e., decision tree SVM. The research paper includes four different sections: Sect. 2 deals with literature survey of both online and offline signatures. Section 3 introduces the proposed work on feature extraction. Last section depicts experimental results and conclusion of paper.

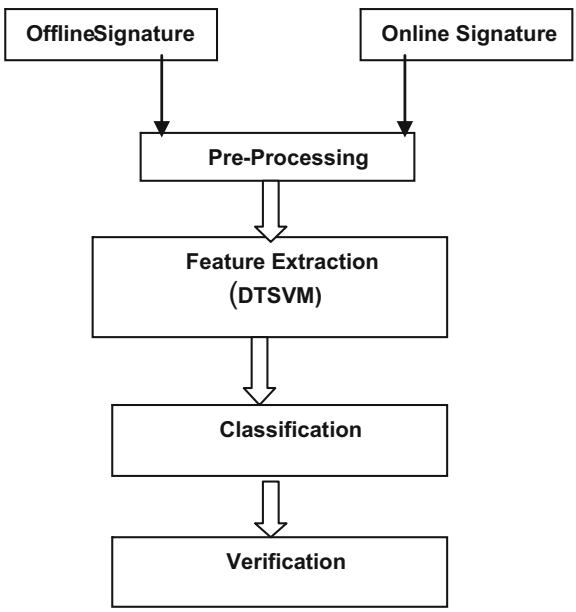


Fig. 1 Block diagram of integrated signature verification

drksujatha23@gmail.com

2 Literature Review

Dakshina Ranjan Kisku et al. (2010) described a technique for offline signatures. The proposed system uses three dissimilar classifiers to identity signature. A novel technique applied on signature image for feature extraction is applied on extracted features that were fused and make a concatenated feature set which is then passed through the three classifiers. The system was tested on more than 5 K handwritten signatures of more than 500 individuals, and the results were efficient than previous researches.

Kruthi et al. (2014) developed a system to identity verification of signature image using Support Vector Machine (SVM) for offline signature. Individual signatures were taken, and set of signatures were collected; then these signatures were preprocessed before extracting the features like CoG, centroid, calculate the number of loops, horizontal profile, vertical profile, normalized area, etc. A hyperplane curve was drawn using the values of features from the SVM classifiers. The developed mechanism is tested on 336 signatures, and 7.16% error was found which was efficient than earlier work.

Özgündüz et al. (2010) presented a system for recognition of offline signatures using global, directional, and grid features. A novel technique was used, i.e., Support Vector Machine (SVM) to verify and classify the signatures and a 0.95% classification ratio was obtained from the proposed algorithm. The performance was compared by backpropagation method as well.

Abdelrahman et al. (2013) described signature verification system especially for offline signatures using support vector machine technique. DTM algorithm was used to bring in line all the signatures for global feature extraction. Various forged and unforged signatures are taken from individual for training, a database of 2 K signatures from approx. 80 individuals, and the performance of approximately 82% is achieved.

Sanjay et al. (2011) proposed an identification and verification system, which was based on the contourlet coefficient for offline and online signatures and Support Vector Machine (SVM), used for classification of the signature images. The signature images were normalized, and then contourlet coefficients were computed on particular extent and direction using contourlet transform in feature extraction. The database of English signatures GPDS-960 is used for experiment and achieved 94% identification rate.

Hanmandlu et al. (2011) presented a system based on Genetic Algorithm-Support Vector Machine (GA-SVM) for online signatures. Different groups were generated for 75 features, and SVM was applied to calculate their performance. The main aim to propose the method is to reduce the computational complexity. The experimental results showed better and accurate performance in terms of accuracy and EER.

Berkay et al. (2012) presented their work on signature's local histogram features. The signatures were divided into HOG and LBP. The global SVM classifier was

drksujatha23@gmail.com

trained different signatures. The experiment done using GPDS-160 signature database achieved a 15.41% EER.

Kumar et al. (2017) described their work on recognition of offline signature. They used certain features of the signature, and Support Vector Machine (SVM) classifier was used to classify and recognize the signature image. The artificial neural networks are trained to match the structural parameters along with the local variations of the signature image. The use of SVM classifier increased the performance by approximately (80%) accuracy.

From the above literature survey, it is clear that previously lots of work have been done on either offline or online verification system using the support vector machine and their various techniques. Here, we proposed our work on integrated signature verification, which is classified using decision tree support vector machine algorithm. The purpose of our research is to increase the accuracy in identification of handwritten as well as digital signature and it also helps to reduce the forgery factor from the signature image. Using decision tree SVM, the overall performance of our system is increased approx 85%.

3 Proposed Work

The proposed algorithm used for accomplishing the integrated signature verification system consists of the following main steps.

3.1 Data Acquisition

The database can be collected from individuals either by sign on a piece of paper or through any digital device. Further, the signatures converted into digital format by scanning. In this paper, both forged and genuine signatures are taken from 100 individual persons to create the own database.

3.2 Preprocessing

Raw signatures are collected from the persons, and further few preprocessing operations are required on (offline as well as online) signatures. That will increase the performance of overall system. Preprocessing is generally done to eliminate the unwanted factor from the signature image. At the time of scanning, certain unwanted pixels are added with the image. In the preprocessing phase, certain operations have been applied to the signature to extract the local, global, and geometric features from the image. Such operations are enhancement, conversions, thinning, smoothing, morphological, etc.

drksujatha23@gmail.com

For geometric normalization, a resize algorithm is used to normalize size. In order to remove the noise and other irrelevant data, various algorithms are used, such as median and mean filters, Canny edge detection, etc. Once the signature is noise free, then that signature is converted into the binary image and outline of the image is obtained. In the online signature image, preprocessing is not the important phase, because the image is taken through the digital device which is already a binary image. However, the noise is required to remove from the signature. The binary image contains either 0's or 1's, where 0's specifies the boundary of the signature and 1's shows the blank area. A threshold value is to set to perform the above value. For the proper and cleaned signature image, we perform the image filling or image smoothing operation, which provide the signature in proper format and help in extracting the more and more features from the image.

3.3 Feature Extraction Method

Feature selection is crucial and important part of signature classification and verification problem. In the proposed work, dissimilar statistical techniques are used to extract feature set [3] separately. Given feature set is extracted from the individual's signature and generate the score. For classification, generated score is fused into our proposed algorithm, i.e., Decision tree support vector machine DTSVM. In our proposed signature verification system, some set of features that are extracted are given below:

S. no.	Features	Description
1	Width of a signature image (W)	The horizontal distance between two points in the signature image
2	Height of a signature image (H)	The vertical distance between two points in the signature image
3	Aspect ratio (AP)	Ratio of width to height of a signature image AP: W/H
4	Horizontal projection (HP)	Counted no. of black pixels of images of horizontal projection (5)
5	Vertical projection (VP)	Counted no. of black pixels of images of signature of vertical projection (5)
6	Area of black pixels (ABp)	Counted no. of black pixels from the binary thinned signatures
7	Normalized area of black pixels (NABp)	By dividing the area of whole signature from the area covered by black pixel of binary, thinned, and skeletonised signature images
8	Center of gravity (CoG)	Obtained from addition of all x, y locations of gray pixels and further divided by the number of pixels calculated
9	Maximum and minimum black pixels (MinBp)	Maximum black pixels are the highest frequency of the black pixel and the minimum is the lowest frequencies of black pixels in the vertical projection, respectively

(continued)

drksujatha23@gmail.com

(continued)

S. no.	Features	Description
10	Global baseline (GB)	It is median of the black pixel distribution and calculated from the two outer most points any projection
11	Upper and lower edge limits (UEL)	It is a difference b/w smoothened and original curves of vertical projection. Above baseline is known as upper edge limit and below baseline is known as lower edge limit
12	Middle zone (MZ)	Distance between upper and lower edge limits is known as middle edge or middle zone

In the next stage, we combine all the features (local & global) and generate a feature set into a feature vector and that feature vector is fused into the decision tree SVM for the image classification and generating matching scores.

4 Decision Tree Algorithm

Step 1: Input the training dataset T = {(x1, y1), (x2, y2),…, (xn, yn)}. Construct the linkage binary tree based on the "best separating principle" (13).

Step 2: Initiate with all classes at the tree top node, find the partition of the corresponding group of classes v/s the rest using twin support vector machine, and also make the partition that generate a horizontal cut in the tree to produce two groups of classes in the complete linkage tree.
Calculate the class centers c_i (i = 1 to N) by

$$ci = \frac{1}{|Xi|}\sum_{x \in Xi} x$$

Distance b/w two different classes through Euclidean distance, i and j = ED_{ij} where (i, j = 1 to N)
ED_{ij} = ‖ci–cj‖ where Xi = set of training data contained in class i
|Xi| = no. of elements included in Xi.

Step 3: Largest value of class Li = min EDij calculated from the largest value of EDij where J = 1,…N, j ≠ I. The largest the value of Li, the farthest the classis

Step 4: Calculate the hyperplane and make the separate class K
K = arg max Li where i = 1… N,
If k exist for plural of I, then next smallest distance Li'. Value of k = arg max Li'

Step 5: For Training all the nodes by TWSVM, Repeat step 2

Step 6: Save the decision tree and return.

drksujatha23@gmail.com

5 Experimental Approach

The efficacy of our proposed work is demonstrated on publicly available MCYT-100 dataset. The dataset contains signatures of 100 individuals. The set of each user contains 25 genuine signatures and 25 forged signatures collected from individual. All the signatures are captured from tablet and pen paper, comprise the (x, y) coordinates, width, angle, inclination, azimuth, height, and other local and global features. For enrolment of signature, we select randomly set of five genuine signatures of an individual. Remaining signature is used for testing of our proposed work.

Number of persons: 100 Number of genuine signatures per person:25 Number of forgeries per person:25 Total no of signatures = 5000(100 × (25 + 25))		
Number of genuine signatures used for training: 5		
Number of forgeries used for training: 5		
FAR	FRR	Accuracy
2.6%	**2.34%**	**87.65%**

6 Result

The comparison table of proposed integrated signature verification system with the DTSVM classification is shown in table below. According to the present results, the value of FAR is lower than the given approaches. As the FAR is lower, it increases the performance of the system. In other words, decision tree SVM is more efficient rather than the linear and nonlinear SVM.

- the value of false acceptance ratio

S. no.	Year	Approach	FAR	FRR	Accuracy
1	Samuel Audet (2006)	SVM	–	<25%	50%
2	2011	SVM (Chinese)	6.94	6.40	93.17
		SVM (Dutch)	3.44	3.86	96.35
3	2013	Virtual SVM	11.00	2.00	13.00
4	Bailing Zhang (2014)	SVM	2.0%	2.5%	99%
5	Rashika Shrivastav (2016)	SVM + DTW-Gabor filter	4.2857%	0%	–
6	R. Kumar (2017)	SVM	–	–	80%
7	Proposed work	Decision tree SVM	∼2.6%	2.34%	∼87.65%

drksujatha23@gmail.com

7 Conclusion

In above research, we present decision tree–support vector machine for multi-classification (a hybrid algorithm for hybrid signature verification system). As we all aware that signatures are unique to identify the human verification, the major issue associated with forgery or duplicacy of signature from the genuine signatures, so the verification of signatures is highly necessary approach. Decision tree SVM is used as a key solution of signature verification by classifying their features for discriminating genuine and forgery signatures. A broad range of experiments were conducted on different datasets to evaluate the performance of the proposed method. The proposed technique, using MCYT dataset, provided significantly better results for skilled forgery; FAR of 2.6% and FRR of 2.34% are obtained.

References

1. Kruthi C (2014) Offline signature verification using support vector machine. In: Signal and image processing (ICSIP), 2014 fifth international conference on IEEE, January 2014. ISBN: 978-0-7695-5100-5
2. Kolhe ST, Pawar SE (2012) Offline signature verification using neural network. Int J Mod Eng Res (IJMER) 2(3):1171–1175. ISSN: 2249-6645
3. Rosten E, Drummond T (2005) Fusing points and lines for high performance tracking. In: Tenth IEEE international conference on computer vision. ICCV 2005, vol 2, pp 1508–1515, October 2005
4. Baltzakis H, Papamarkos N (2001) A new signature verification technique based on a two-stage neural network classifier, Pergamon, pp 95–103
5. Abdelrahman A, Abdallah A (2013) Signature verification system based on support vector machine classifier. In: The international Arab conference on information technology (ACIT'2013)
6. Yadav M, Kumar A, Patnaik T, Kumar B (2013) A survey on offline signature verification. Int J Eng Innov Technol (IJEIT) 2(7)
7. Nguyen V, Blumenstein M, Muthukkumarasamy V, Leedham G (2007) Off-line signature verification using enhanced modified direction features in conjunction with neural classifiers and support vector machines. In: Proceedings of the 9th international conference on document analysis and recognition, vol 2, pp 734–738, September 2007
8. Papunzarkos N, Baltzakis H (1977) Offline signature verification using multiple neural network classification structures. In: IEEE conferences 1977
9. Ghandali S, Moghaddam ME (2008) A method for off-line Persian signature identification and verification using DWT and image fusion. In: IEEE international symposium on signal processing and information technology, pp 315–319
10. Coetzer J, Herbst BM, du Preez JA (2004) Offline signature verification using the discrete radon transform and a hidden markov model. EURASIP J Appl Signal Process 559–571:2004
11. Singhal P, Kumar R (2017) Signature verification using support vector machine (SVM). Int J Sci Res Manag 5(5):5327–5330
12. Malik MI, Liwicki M, Dengel A, Uchida S, Frinken V (2014) Automatic signature stability analysis and verification using local features. In: 14th international conference on frontiers in handwriting recognition, pp 621–626. IEEE

drksujatha23@gmail.com

13. Shao Y, Deng N, Chen W, Wang Z (2013) Improved generalized eigenvalue proximal support vector machine. IEEE Signal Process Lett 20(3):213–216
14. Xu Y, Zhong P, Wang L (2010) Support vector machine-based embedded approach feature selection algorithm. J Inf Comput Sci 7(5):1155–1163

drksujatha23@gmail.com

Securing Web Access—DNS Threats and Remedies



Anchal Sehgal and Abhishek Dixit

Abstract The Domain Name System (DNS) is a distributed approach of mapping the domain name attributes with their respective IP addresses. The DNS has inadvertently been an attack target since its inception. The attack vectors have subsequently given rise to the amendments in protocols and various other techniques that defer or mitigate the risk factors of the same. In this paper, an analytical review of the vulnerabilities and threats of the DNS and its various security implications has been discussed. DNS Security Extension (DNSSEC), a public key cryptographic approach proposed by the Internet Engineering Task Force (IETF) to protect the integrity of the DNS records, has further been presented. Further, in this paper, the DNSSEC architecture, usage, and operational security flaws which subsequently might compromise the integrity of the DNS are also presented.

1 Introduction

The advent of DNS in 1983 facilitated the memorization of web addresses using some alphanumeric string instead of a long 32-bit binary address or a 12-digit numeric address. The DNS system is responsible for mapping a domain name like www.example.com to an Internet Protocol (IP) address. Pointing the request searching for specific domain to its IP address is carried out by DNS server.

The DNS system, as a trade-off, while helping with the resolution of websites, also brought some loopholes with it. By the time of design of DNS, scalability being a prime concern, security for malicious behavior was not included. The attacks on the credibility and availability of the websites have also made the DNS security a prime concern for the website administrators [1, 2].

A. Sehgal
Arya Institute of Engineering and Technology, Jaipur, India
e-mail: anchal.6455@gmail.com

A. Dixit (✉)
Jaipur Engineering College and Research Centre, Jaipur, India
e-mail: abhishekdixit.cse@jecrc.ac.in; abhishekdixit2606@gmail.com

© Springer Nature Singapore Pte Ltd. 2019

V. S. Rathore et al. (eds.), *Emerging Trends in Expert Applications and Security*,
Advances in Intelligent Systems and Computing 841,
https://doi.org/10.1007/978-981-13-2285-3_40

drksujatha23@gmail.com

DNSSEC by IETF attempts to add some security to the existing DNS architecture by maintaining its backward compatibility. Such mechanism uses public key cryptography, thereby maintaining DNS records integrity and adding authenticity to it.

2 DNS Infrastructure

The URL typed in browser is tried to be resolved into its IP addresses by the stub resolver of the operating system. The local system cache is searched by the stub to map the request to its IP addresses. Upon nonavailability of any such record, the request is forwarded to the Recursive Name Server (RNS) managed by Internet Service Provider (ISP). The RNS at ISP searches its cache records for respective IP address, upon nonavailability of which it recursively or iteratively forwards it to root Name Server (NS) which resolves the request and dispenses the IP address of NS managed by domain.

2.1 DNS Hierarchy

The naming convention of any domain is of the form www.example.com, where each of these words represents different levels of hierarchy in the DNS architecture. The hierarchy of their respective servers corresponds to different zones in a domain and the zones are separated by ".", e.g., www.example.com. Here, "com" is the TLD and is at the highest level of this hierarchy which is controlled by DNS root zone. TLD records points to DNS server at the lower level usually at NS. This NS is also called as Authoritative Name Server (ANS) of that domain [3]. The records of the ANS subsequently point to its mail servers, subdomains, A-records, etc.

2.2 Resource Records and Resolvers

The records in a DNS server are name and IP address associations and are called as Resource Records (RR). Any particular domain name can have multiple records (as shown in Fig. 1a and b) with different priorities assigned for any particular set [4]. Resolvers are library extensions present in DNS server to resolve IP address of any domain.

drksujatha23@gmail.com



(a)

A	a host address
NS	an authoritative name server
MX	mail exchange
TXT	text strings
RP	for Responsible Person
X25	for X.25 PSDN address
ISDN	for ISDN address
AAAA	IP6 Address
LOC	Location Information
CNAME	the canonical name for an alias
SOA	marks the start of a zone of authority

(b)

Fig. 1 **a** Snapshot of DNS resource records. **b** Different DNS resource records

2.3 DNS Packet Structure

All DNS packets follow the structure shown in Fig. 2[5].

Header: Header describes the packet type and the fields that are in the packet. The request to the DNS server and the response from it follow the same format as shown in Fig. 3. The description of each of these fields is described as follows:

ID: an identifier of 16 bits.

QR: specifies whether the packet is query or a response packet. This is a 1-bit field.

OPCODE: It specifies the kind of query. This is a 1-bit field.

AA: Authoritative Answer—used only in response specifying whether the response is from an authoritative name server. This is a 1-bit filed.

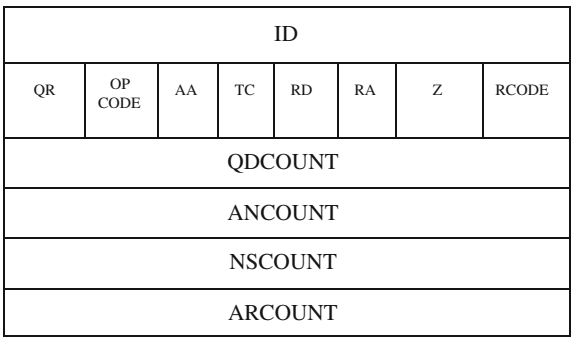
TC: Truncation—specifies whether the message is truncated or not.

Fig. 2 DNS packet structure

Header
Question
Answer
Authority
Additional

drksujatha23@gmail.com

Fig. 3 DNS packet header



RD and RA: abbreviates for recursion desired and recursion available, respectively. These are 1-bit fields. The former directs whether to forward the query recursively or not. And the latter is used in response to specify whether recursive query is available or not.

Z: reserved for future use.

RCODE: Response Code—This value is of 4 bits and is used to respond whether there was any failure, error, refused, or no error.

QDCOUNT, ANCOUNT, NSCOUNT, and ARCOUNT: These are 16-bit values. They are used to represent questions count, answers count, number of name servers count, and number of additional records count, respectively.

DNS Question: DNS Question follows the format defined as follows:

QNAME: This field specifies the domain name, which constitutes a series of labels, where every label further contains octet length and the total number of octets.

QTYPE and QCLASS: These fields represent the query type and query class, respectively, by two octet codes.

DNS Answer: DNS Answer follows the format defined as follows:

NAME: This field is same as QNAME present in DNS Question format, means it gives us the domain name which was queried.

CLASS: This field consists of two octets specifying the data class present in the RDATA field.

TTL: Time in which the results can be cached. Represented in seconds.

RDATA: This field specifies the response data. The format of this field is more dependent on the filed "TYPE" in the following ways:

- if "TYPE" field is 0x0001 for A-records, then this represents that this is the IP address having four octets.
- if "TYPE" field is 0x0005 for CNAMEs, then this specifies the name of an alias.
- if "TYPE" field is 0x0002 for name server, then this specifies the server name.
- if "TYPE" field is 0x000f for mail server, then it follows two entities, namely, PREFERENCE and EXCHANGE. The former is a 16-bit integer defining the

drksujatha23@gmail.com

preference of mail server, whereas the latter is a domain name with the same format as that of QNAME.

DNS Authority and Additional: These fields in DNS packet follow the same format as DNS Answer. Their order of appearance in packet determines which field they belong to.

3 DNS Vulnerabilities and Attacks

DNS has been a major area of interest for the Black Hat community since the time of its inception. The flaws in design have a major hand in the attacks that have been prevalent over time. Some of the attacks suffered by DNS have been identified in literature.

3.1 DNS Cache-Based Attacks

The time taken by the RNS to resolve queries is considerably large in terms of performance. To reduce the time and enhance the speed of generating response, local cache memory is used at servers. The requests that are frequently made are stored in cache for faster response. The TTL value of RR is the validity of the record in cache memory. In other words, when the TTL expires, the higher level server is queried again instead of sending the RR present in the cache.

Cache Poisoning

Cache poisoning can be achieved by spoofing the IP address in the response DNS packet. A DNS packet can easily be intercepted by an attacker and can be changed to the attacker's chosen IP addresses. The RDATA of a DNS packet can be a carrier for a DNS name, which can be exploited by the attacker to inject false data into a cache using additional data section of response packet. Arbitrary DNS names can also be introduced that can inject false information into the victim's cache.

3.2 Man-in-Middle Attacks

DNS does not provide any way for origin authenticity or data integrity that it receives. Originator's source IP address is the only way of providing authenticity that can be easily spoofed by an attacker. This flaw in the design is exploited by attackers to create forges response packets. The absence of any encryption or digital

drksujatha23@gmail.com

signature makes it easy for an attacker to modify the DNS response packets. The packets can be easily captured by an attacker by intercepting the propagation channel. The absence of integrity and authentication mechanisms enables changes to go undetected [6].

3.3 Denial-of-Service Attacks

The Denial-of-Service (DOS) attacks for the DNS are of distributed nature and can have a significant downtime effect over the Internet community on a global scale. The DOS attacks are generally targeted at the root zone, as their downtime would ensure the proper downtime of all sites of that particular zone. There have been consistent attacks of this nature over time [6].

DDOS Amplification Attacks

The DNS amplification attacks aim at increasing the amount of bandwidth to target a victim. A DNS request is of 64 bytes which generates a response of 3223 bytes. This attains an amplification of over 50. An attacker who has the control of a botnet of say 100 Mbps would send request packets and hence would generate response of 50X size and hence would require a bandwidth of 5 Gbps for response. The queuing in of requests would subsequently enqueue response packets and shall generate large traffic that will consume the full bandwidth and hence cripple the network.

3.4 DNS Rebinding Attacks

This attack is one of the most prominent attacks in DNS and is initiated by a small advertisement luring a victim to visit the attacker's website.

3.5 DNS Tunneling

DNS tunneling is used to transfer data as a payload of DNS packets. It can be used to exfiltrate data or can be used in tunneling any IP traffic. Many tools are available to implement DNS tunneling. This can be used to circumvent firewall or any packet filtering services. An example would be a request packet asking for the C-Name of THISISTHEPAYLOAD.example.com. The server can give a response as THISISMYRESPONSE.THISISTHEPAYLOAD.example.com. One can send up to 255 characters including the subdomains and each subdomain of up to 63 characters using such encoding.

drksujatha23@gmail.com

4 Domain Name System Security Extension (DNSSEC)

DNSSEC is a set of security extensions introduced by IETF. This adds security to the DNS by providing data integrity and origin authentication. This also provides an authenticated non-existence of DNS data, provided by ANS. DNSSEC application reduces DNS attacks substantially; however, few attacks such as DOS cannot be curbed by it.

4.1 Infrastructure

DNSSEC is an extension of the DNS protocol. Some amendments have been made to the DNS architecture to add security to it. This ensures its maintenance of backward compatibility.

DNSSEC Resource Records

The DNSSEC has introduced a few new record types to be used for security purposes. The added record types are as follows:

- Resource Record Signature (RRSIG)—It contains the digital signature of any record set. And the resolvers verify this with the stored public key.
- DNS Key—This is the public key used to verify RRSIG records.
- Delegation Signer (DS)—Name of a delegated zone.
- Next Secure (NSEC)—It lists the records that exist for that particular record and also contains a link to the next record.
- Next Secure3 (NSEC3)—The links to next records are hashed and are in sorted order. These records verify the non-existence of records.
- NSEC3PARAM—This is used by ANS to determine which NSEC3 record is to be sent for non-existent record's response [4].

5 DNSSEC in Mitigating Threats

By the virtue of data integrity and origin authentication property of DNSSEC, it is possible to mitigate and prevent cache-based and man-in-middle attacks of DNS.

As already discussed, cache-based attacks can also be initiated by man-in-middle attack. An attacker can intercept the response packet and easily alter the ANS IP address to point to any IP address of his choice. With DNSSEC, RR's integrity can be verified by RNS or resolver using its digital signature present in the DNSKEY record of DNS response packet [6].

drksujatha23@gmail.com

6 Other Preventive Measures Against Various DNS Threats

6.1 DNS Rebinding Attack

This attack initiates the execution of a malicious client-side script by establishing trust through a legitimate site and then transferring the web control to a malicious site by forcing to recheck the A-records upon reduction of the cache TTL. Unfortunately, DNSSEC does not have the feature of detecting or preventing this attack. Alternatively, defense for such an attack can be arranged through third-party plug-ins, packet filtering policies of firewalls, or by exclusive RNS policies.

6.2 DNS Tunneling

As discussed above, DNS tunneling involves the encapsulation of data within DNS queries and responses using various encoding techniques. The utilization of DNS resolution helps in tunneling the traffic over other systems through Internet. The basic prevention technique for DNS tunneling is the continuous monitoring of DNS traffic in a real-time environment.

7 Conclusion

Domain name system has come a long way with its existence in resolving domain names with their existing IP addresses. Lately due to some threats and attacks, DNSSEC which was proposed by IETF as an extension to the existing DNS infrastructure helped as a countermeasure to the threats and vulnerabilities. DNSSEC maintains security by providing origin authentication and data integrity for the resource records of a domain. Various proposals have been made to encumber such threats and attacks, which predominantly have been helpful to tackle this to a large extent.

References

1. Kovaes E (2015) Google Vietnam targeted in DNS hijacking attack. Retrieved from www.securityweek.cim/google-vietnam-targeted-dns-hijacking-attack
2. Russon M-A (2015) Google Vietnam domain name briefly hacked and hijacked by Lizrd Squad. Retrieved from www.ibtimes.co.uk/google-vietnam-domain-name-briefly-hacked-hijacked-by-lizard-squad-1489293

drksujatha23@gmail.com

3. Mealling M, Daniel R (2000) The naming authority pointer (NAPTR) DNS Resource Record. RFC-2915
4. Arends R, Austein R, Larson M, Massey D, Rose S (2005) Resource records for the DNS security extensions. RFC-4034
5. Mockapetris P (2004) Domain names implementation and specification. RFC-1035
6. Ariyapperuma S, Mitchell CJ (2007) Security vulnerabilities in DNS and DNSSEC. In: ARES-the second international conference on availability, reliability and security, IEEE Computer Society

drksujatha23@gmail.com

P2S_DLB: Pluggable to Scheduler Dynamic Load Balancing Algorithm for Distributed Computing Environment



Devendra Thakor and Bankim Patel

Abstract Unbalanced load is one of the major issues of distributed system which leads to worst utilization of available resources. All existing dynamic load balancing algorithms (DLBA) start balancing activities after the system becomes unbalanced. It is more appropriate to design load balancing algorithm that can plug into scheduling algorithm and helps scheduling algorithm to schedule incoming jobs in such way that system remains in balanced state without increasing unnecessary system overhead. In this paper, we have designed a pluggable to scheduler DLBA which can schedule incoming jobs by considering current load status of clusters. The designed algorithm is plugged with priority scheduling algorithm. The experimental results show that designed algorithm improves cluster utilization over scheduling and predication based algorithms.

Keywords Dynamic load balancing · Scheduling · Distributed system
Cluster

1 Introduction

The advancement in computing and communication technology increases the usage of distributed computing environment. Distributed system provides platform to access geographically scattered resources by sharing jobs between computational nodes. The geographically scattered computational resources of distributed system are congregated into virtual groups known as cluster [1]. The users of distributed system generate jobs which are assigned to different clusters for the execution. The biggest issue of clustered environment is the uneven utilization of clusters in the system [1]. It has been noted in the literature that gradually some of the clusters

D. Thakor (✉) · B. Patel
Uka Tarsadia University, Bardoli, Surat, India
e-mail: devendra.thakor@utu.ac.in; devendrathakor@gmail.com

B. Patel
e-mail: bankim_patel@srimca.edu.in

© Springer Nature Singapore Pte Ltd. 2019

V. S. Rathore et al. (eds.), *Emerging Trends in Expert Applications and Security*,
Advances in Intelligent Systems and Computing 841,
https://doi.org/10.1007/978-981-13-2285-3_41

drksujatha23@gmail.com

become overloaded while the other clusters remain medium-loaded or under-loaded in distributed system [2, 3]. It is the problem of unbalanced load between clusters which reduces the overall system performance as resources are not utilized properly. The problem can be solved by applying dynamic load balancing techniques.

Load balancing is the process of harmonizing load between clusters of distributed system [1–5]. It aims to optimize resource utilization, maximize throughput, minimize response time, and avoid overload of any single resource. Load balancing algorithms improve the performance of distributed system by redistributing load between under-loaded, medium-loaded, and overloaded clusters. The load balancing algorithms can be categorized in two types, static and dynamic [2–5]. The static algorithm does load balancing by considering information available at compile time; while dynamic algorithm balances the load between clusters on the basis of run time information [2–5]. The study of existing DLBA plays important role to highlight the limitations and justifies the need of designing an efficient DLBA [6, 7].

Existing DLBA starts balancing activities after system becomes unbalanced. The better approach is to start balancing activities from the beginning by running DLBA in parallel with job scheduling algorithm. The scheduling algorithm schedules incoming job in a well-defined manner without considering current status of clusters. If the scheduling decision is to be taken on the basis of current status of clusters then the problem of unbalanced load between clusters can be resolved. Therefore, the designing of new DLBA is fairly desirable, which can plug to any scheduling algorithm to select best suitable cluster by providing current load status of clusters. The approach is not addressed well in literature [2–7].

The first objective of the study is to design DLBA which can run with scheduling algorithm and start load balancing activity at the time of job allocation to clusters. The designed DLBA should be generalized so that it can be applied to any scheduling algorithm. Second objective is to apply designed DLBA to selected scheduling algorithm. The priority scheduling algorithm is selected for testing and checking performance. Third objective is to do severe comparative analysis of the experimental results of proposed algorithm with priority scheduling algorithm and prediction based algorithm.

2 Related Work

The issue of unbalanced load between clusters is addressed well in literature. The researchers came up with many approaches for balancing load in distributed system which can be found in the literature [6–13].

In [6], authors explore and explain the essential components for designing an effective DLBA for distributed computing. An adaptive decentralized sender-initiated load balancing algorithm that utilizes the load estimation approach is presented in [7]. The algorithm is adaptive as it can estimate different types of strongly influencing system parameters. A dynamic threshold based scheduling and

drksujatha23@gmail.com

load balancing algorithm is proposed in [8]. The algorithm dynamically updates the threshold values with respect to runtime environment changes in resource workload. It considers workload of resources as load index instead of number of tasks in queue because weights of tasks are different.

In [9], the authors propose the prediction-based dynamic load balancing technique for heterogeneous clusters. The proposed technique predicts three different types of resources like processor, memory, and IO requirements of incoming jobs. The DLBA which can predict the neighbors load information is proposed in [10]. The algorithm balances the load on the basis of predicted load information of neighbors node. In [11], the authors propose a protocol for dynamic load balancing in grid. The proposed protocol is scalable and has low message and time complexities. The protocol primarily focuses on when and where the load should be transferred but the authors have not addressed the problem of how the load should be transferred. A hybrid DLBA for heterogeneous environments is proposed in [12]. They proposed two new algorithms for super node selection in a cluster and analyzed the performance of algorithms under different cluster configurations, load scenarios, and network topologies.

These existing dynamic load balancing algorithms attempts to enhance the performance of heterogeneous clustered environment by improving different parameters. The algorithms [9], tries to predict the future requirements of incoming jobs and [10] tries to predict the neighbors load status. Our previously proposed prediction based DLBA (PDLB) [13] predicts the future status of clusters. The predicted status of clusters is playing vital role to balances load between the clusters. The experimental results show that prediction approach improved the resource utilization but still there is a scope of improvement. It indicates that prediction is not perfect solution of unbalanced load issue in clustered environment.

3 Pluggable to Scheduler Dynamic Load Balancing Algorithm (P2S_DLB)

The proposed DLBA is designed on the basis of an idea of doing load balancing activity at the time of job scheduling. The algorithm takes current status of computing resources in clusters and incoming job as an input and find out best suitable cluster for incoming job execution as an output. The clusters having higher number of free resources are selected for incoming job execution. In algorithm, n and m indicate total number of nodes and clusters, while PE means processing elements.

The proposed algorithm has total four stages, in first stage it tries to achieve more than 50% utilization of available resources in each and every clusters. The step 3 finds out cluster having more than 50% free resources and step 4 schedules incoming job to find out clusters for execution, if the cluster is suitable to execute incoming job. In, second stage an algorithm going one step ahead and tries to

drksujatha23@gmail.com

achieve more than 66% utilization of available computational resources. The step 5 and 6 are executes

Algorithm 1 Pluggable to scheduler dynamic load balancing algorithm (P2S_DLB)

```
1: for i =1 to n do
2:     for  j =1 to m do
3:         if Num_Free_PE_j > Num_PE_j/2 then
4:             if Num_Running_PE_j ≥ Num_Required_PE_i then
5:                 Schedule i^th job to j^th cluster for execution
6:         else if Num_Free_PE_j > Num_PE_j/3 then
7:             if Num_Running_PE_j ≥ Num_Required_PE_i then
8:                 Schedule i^th job to j^th cluster for execution
9:         else if Num_Free_PE_j > Num_PE_j/4 then
10:            if Num_Running_PE_j ≥ Num_Required_PE_i then
11:                Schedule i^th job to j^th cluster for execution
12:        else if Num_Free_PE_j > Num_PE_j/5 then
13:            if Num_Running_PE_j ≥ Num_Required_PE_i then
14:                Schedule i^th job to j^th cluster for execution
15:        else
16:            if Num_Running_PE_j ≥ Num_Required_PE_i then
17:                Schedule i^th job to j^th cluster for execution
18:    update load status of C_j
```

when the resource utilization of all clusters is more than 50%. The step 5 find out clusters having more than 33% free resources and step 6 schedules incoming job to the find out clusters if clusters has suitable resources. In, third and fourth stages the algorithm tries to achieve 75 and 80% resource utilization, respectively. The steps 7–8 and 9–10 executes when cluster utilization reaches 75% and 80%, respectively.

The incoming jobs are scheduled in such a way that indirectly it balances the load between clusters. The algorithm achieves better resource utilization and increases the level of load balancing step by step. It dynamically balances load by scheduling incoming job to specific cluster, considering current load status of that cluster. The best part of proposed algorithm is, it balances load without any job migration from one cluster to other cluster.

4 Implementation Scenario and Input Dataset

In order to demonstrate improvement in load balancing, Windows 10 on an Intel Core i3–370 M processor with 3 GB of RAM and 320 GB of hard disk is used during experiments in ALEA version 2 [14] with JDK 1.8 and JRE 1.8. ALEA is Gridsim [15] based job scheduling simulator. The GridSim is java-based simulation tool that provide facilities to model and simulate the entities like users, heterogeneous resources, resource load balancers, and applications.

drksujatha23@gmail.com

The real-time dataset from the Gaia cluster log is taken for testing algorithms. The Gaia cluster is one of the four clusters operated by the University of Luxembourg HPC Center (ULHPC) initially released in 2011 [16, 17]. The Gaia is a heterogeneous cluster that has been upgraded several times. The selected data set contains three months of data from May to August, 2014 in which total 51,987 jobs are created. The Gaia cluster is collection of total eight heterogeneous clusters, which have total 151 nodes manufactured by Bull and Dell and total 2004 cores for job execution [16, 17].

5 Experimental Results and Analysis

The performance P2S_DLB algorithm is evaluated and compare against the traditional approach of priority scheduling (PS) and our previously proposed prediction based dynamic load balancing (PDLB) algorithm. The percentage of cluster utilization is considered as performance evaluation parameter.

Table 1, show the cluster utilization and average cluster utilization. The results show that average cluster utilization is increased by almost 15 and 9% in P2S_DLB compared to PS and PDLB, respectively. The individual cluster utilization is improved in seven clusters out of eight clusters in proposed approach. In the existing approach the utilization of first cluster which has the highest amount of processing power is 70.67%. It indicates that most of the incoming jobs are executed on first cluster. The utilization of other clusters increased in PDLB as the utilization of first cluster reduced by 6.81%. The utilization of first cluster further reduces by 12.73% in P2S_DLB which increases the utilization of under-utilized clusters. The utilization of first cluster is 70.67% in PS and 63.86% in PDLB, on the other hand fifth cluster utilizes only 8.58% in PS and 19.05% in PDLB which show unbalanced load and uneven resource utilization.

Table 1 Cluster utilization in percentage

Cluster Id	Cluster name	PS	PDLB	P2S_DLB
1	gaia-[1–60]	70.67	63.86	51.13
2	gaia-[61–62]	10.01	19.71	38.04
3	gaia-[63–72]	13.91	24.63	39.38
4	gaia-73	16.28	27.38	37.54
5	gaia-74	8.58	19.05	38.54
6	gaia-[75–79]	10.05	21.42	35.37
7	gaia-[80–119]	42.68	41.30	45.02
8	gaia-[120–151]	33.19	35.76	40.25
	Average utilization	25.67	31.64	40.66

drksujatha23@gmail.com

The utilization of second to sixth clusters is very less in existing approach. It shows that balancing is required among the clusters. The problem is resolved in prediction approach as the PDLB improves utilization of second to sixth clusters by 10% which increases the average utilization by 6%. The P2S_DLB algorithm performs better compared to PDLB and increases the utilization of second to six clusters by 15% which improves the average utilization by 9%. The less utilization of second and fifth clusters in PS, which has least processors show that PS under utilizes the clusters having less processing unites. The PDLB improves the utilization of second and fifth clusters but compared to the other clusters they remains at last positions. The P2S_DLB overcomes this problem by utilizing the second cluster 38.04% and fifth cluster 38.54% which is close to average utilization and higher than the other clusters. It show that the performance of P2S_DLB is not depends on number of processing unites in clusters and thus P2S_DLB is best suited algorithm for heterogeneous clusters.

Figures 1, 2 and 3 show the cluster usage on day-to-day basis in PS, PDLB, and P2S_DLB respectively. In Figs. 1, 2 and 3 X-axis represents number of days and Y-axis represents the clusters in descending order of CPUs. Figures 1, 2 and 3 show the total eight rows separated by thin white line, represents individual cluster as dataset has eight clusters. In figures, rows are painted for 90 days because the selected dataset for simulation is of three months. The green, yellow, and red colors represent less, medium, and high clusters utilization respectively.

Figure 1 show the performance of priority scheduling algorithm. The first row is painted with red color in most of the portion which indicates that first cluster having highest CPUs remains busy for all days. The second row is painted with red color in half of the portion which indicates cluster having second highest CPUs remains busy for half of the days. The remaining rows are painted with green color in most of the portion, means clusters having less CPUs remains free on most of the days. It shows the PS algorithm is unable to balance the load between clusters.

Figure 2 show the performance of PDLB algorithm where, all rows are painted with red and yellow colors for the period of 20 days, from fifth day to twenty fifth

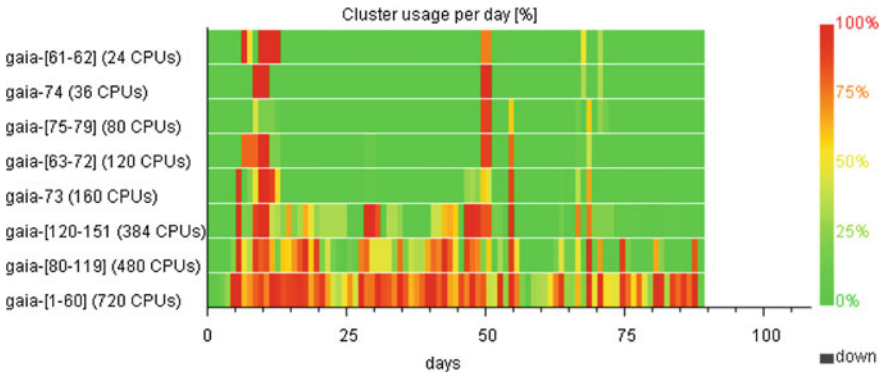


Fig. 1 Cluster utilization in PS

drksujatha23@gmail.com

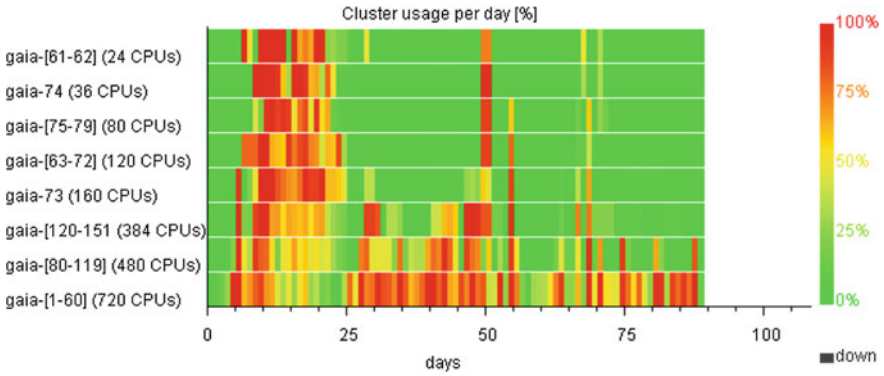


Fig. 2 Cluster utilization in PDLB

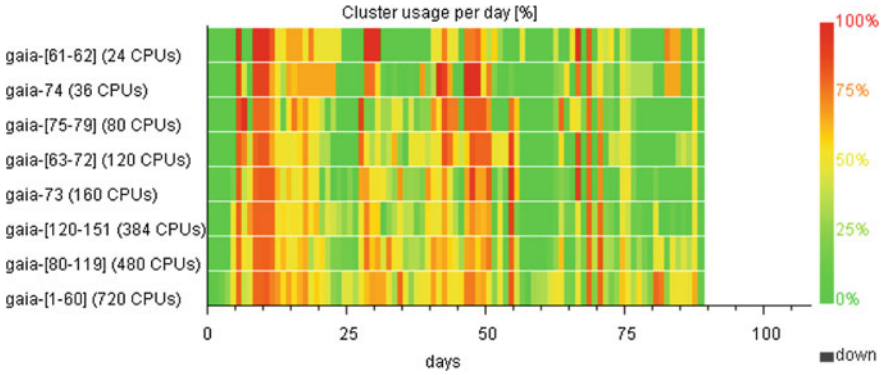


Fig. 3 Cluster utilization in P2S_DLB

day in which highest number of jobs came for an execution. It indicates PDLB can predict the loaded situation of clusters and scheduled incoming jobs to underloaded clusters which increased the overall resource utilization. The limitation of PS algorithm overcome up to some extend in PDLB algorithm. The PDLB balanced the load better compare to PS for first 35 days, but it performed same as the PS for remaining 55 days, which indicates scope of improvement in the PDLB.

Figure 3 shows the performance of P2S_DLB algorithm here, most of the portions of all eight rows are painted with yellow or light red colors which illustrates all the eight clusters remains medium-loaded for most of the days. The frequency of clusters remains in under-loaded condition is less in P2S_DLB as the green painted portion is less, which was high in case of PS and PDLB algorithms. The red painted portion is very less means clusters hardly go in an overloaded condition. Figure 3 shows that the proposed P2S_DLB algorithm performed best in terms of load balancing and optimal resource utilization.

drksujatha23@gmail.com

6 Conclusion and Future Scope

A novel idea of balancing load between resources at the time of scheduling is suggested in paper. The DLBA P2S_DLB is designed which can plug to any scheduling algorithm in distributed system. The algorithm finds clusters with higher numbers of free resources and schedules incoming jobs to that clusters. It increases the average cluster utilization step by step in the system. The P2S_DLB algorithm is plugged to traditional priority based scheduling algorithm. The PS, PDLB, and P2S_DLB algorithms are simulated on real-time dataset and real-time cluster configuration using ALEA simulator. Each day cluster utilization and average cluster utilization are calculated as an evaluation parameters. Results show that P2S_DLB gives best performance compare to PS and PDLB algorithms. The proposed algorithm increases the average cluster utilization of the clusters. As a future scope, designed algorithm can be plugged with other well-known scheduling algorithms for checking stability. The combination of PDLB and P2S_DLB algorithms can be proposed which may further increase average cluster utilization in the system.

References

1. Buyya R (1999) High performance cluster computing: architecture and systems, volume i. Prentice Hall, Upper Saddle River, NJ, USA
2. Sharma S, Singh S, Sharma M (2008) Performance analysis of load balancing algorithms. Int J Comput Electr Autom Control Inf Eng 2(2):367–370
3. El-Zoghdy SF, Kameda H, Li J (2002) A performance comparison of dynamic vs. static load balancing policies in a mainframe-personal computer network model. In: Proceedings of the 39th IEEE conference on decision and control (Cat. No.00CH37187), Sydney, NSW, 2000, vol.2, pp 1415–1420
4. Kameda H, Li J, Kim C, Zhang Y (2012) Optimal load balancing in distributed computer systems. Telecommunication networks and computer systems. Springer, London
5. Rajguru AA, Apte S (2012) A comparative performance analysis of load balancing algorithms in distributed system using qualitative parameters. Int J Recent Technol Eng 1(3)
6. Mehta MA, Jinwala DC (2012) Analysis of significant components for designing an effective dynamic load balancing algorithm in distributed systems. In: 2012 Third international conference on Intelligent Systems Modelling and Simulation, pp 531–536
7. Shah R, Veeravalli B, Misra M (2007) On the design of adaptive and decentralized load balancing algorithms with load estimation for computational grid environments. IEEE Trans Parallel Distrib Syst 18(12):1675–1686
8. El-Zoghdy SF, Elnashar AI (2015) A threshold-based load balancing algorithm for grid comput-ing systems. J High Speed Netw 21(4):237–257
9. Chandra PK, Sahoo B (2010) Prediction based dynamic load balancing techniques in heterogeneous clusters. In: Proceedings of international conference on computer science and technology, pp 189–192
10. Lim JW, Hoong PK, Yeoh ET (2012) Neighbor's load prediction for dynamic load balancing in a distributed computational environment. In: TENCON 2012 IEEE Region 10 Conference, Cebu, pp 1–6
11. Payli RU, Erciyes K, Dagdeviren O (2011) Cluster-based load balancing algorithms for grids. CoRR. abs/1110.1991

drksujatha23@gmail.com

12. Mehta M, Jinwala D (2011) A hybrid dynamic load balancing algorithm for heterogeneous environments. In: Proceedings of international conference grid computing and applications, pp 61–65
13. Thakor D, Patel B (2017) Pdlb: an effective prediction based dynamic load balancing algorithm for clustered heterogeneous computational environment. In: 5th international conference on advanced computing, networking, and informatics
14. Klusa´cekˇ D, Rudova´ H (2010) Alea 2: job scheduling simulator. In: Proceedings of the 3rd international conference on simulation tools and techniques. SIMUTools '10, Belgium, vol 61, pp 1–10
15. Buyya R, Murshed M (2002) Gridsim: A toolkit for the modeling and simulation of distributed resource management and scheduling for grid computing. Concurr Comput Pract Exp 14(13–15):1175–1220
16. Emeras J Parallel workloads archive university of luxemburg gaia cluster. http://www.cs.huji.ac.il/labs/parallel/workload/lunilugaia/. Accessed 15 Feb 2016
17. ULHPC: The gaia cluster hpc at university of luxemburg. https://hpc.uni.lu/systems/gaia/. Accessed 15 Feb 2016

drksujatha23@gmail.com

Parkinson Disease Prediction Using Machine Learning Algorithm



Richa Mathur, Vibhakar Pathak and Devesh Bandil

Abstract Parkinson disease, the second most common neurological disorder that causes significant disability, reduces the quality of life and has no cure. Approximately, 90% affected people with Parkinson have speech disorders. The medical dataset contains heterogeneous data in the form of text, numbers, and images that can be mined. Big Data has the potential to give valuable information after processing that can be discovered through deep analysis and efficient processing of data by decision-makers. Data mining is the process of selecting, extracting, and modeling the unknown hidden patterns from large datasets. Machine learning algorithm (MLA) can be used for early detection of disease to increase the chances of elderly people's lifespan and improved lifestyle with Parkinson. In this paper, we use various MLAs that can help in improving the performance of datasets and play a vital role in making the early prediction of disease at right time. After comparison of these algorithms, we choose the most effective one in terms of accuracy. From our experimental results, it is analyzed that the accuracy obtained from the combined effect of KNN algorithm with ANN is better as compared to other algorithms.

Keywords Parkinson disease · Predictive analytics · Voice datasets
SVM · KNN · ANN

R. Mathur (✉)
Department of CS & Applications, SGVU, Jaipur, Rajasthan, India
e-mail: richa0058@gmail.com

V. Pathak
CS/IT, Arya College of Engineering and IT, Jaipur, Rajasthan, India
e-mail: vibhakarp@rediffmail.com

D. Bandil
Computer Applications, Suresh Gyan Vihar University, Jaipur, Rajasthan, India
e-mail: mcagwailor@gmail.com

© Springer Nature Singapore Pte Ltd. 2019

V. S. Rathore et al. (eds.), *Emerging Trends in Expert Applications and Security*,
Advances in Intelligent Systems and Computing 841,
https://doi.org/10.1007/978-981-13-2285-3_42

drksujatha23@gmail.com

1 Introduction

Parkinson disease (PD), a neurodegenerative disorder of CNS system, also produces movement disorder. In the primary stage when nerve cells or neurons of the brain become impaired, people may begin to notice symptoms like tremor, stiffness in limb or trunk of the body, movement issues, or impaired balance. With the progression of the disease, people may have difficulty in walking, talking, or completing other simple tasks. PD has no cure but several treatments are known to provide relief from the symptoms. PD affects brain cells that produce Dopamine, a neurotransmitter, which is responsible for coordinate and control muscle activity [1].

Typically, PD occurs in people over the age of 60, among which 1% of people are affected. It is called as young onset PD when it is seen in the people before age 50 [2]. According to estimates, PD affected 6.2 million people and about 117400 deaths globally in 2015 [3].

Nowadays, data is becoming more valuable but how to handle data and finding hidden facts from it is more important. The term "Big Data" describes a large amount of datasets that are so complex that it is not possible to process them via conventional methods and technologies. To extract valuable insights from such varied and rapid growing datasets, various tools and techniques of Big Data analytics can be used that may lead to better decision-making and strategic planning.

The rising population generates a large amount of data related to patients clinical and laboratory tests. With this dataset, doctors can detect and diagnose the disease at their early stages. Early prediction of motor symptoms of PD can get a proper treatment at right time to a patient.

2 Related Work

Shamli and Sathiyabhama [4] proposed multi-classifier system, i.e., based on Big Data analytics to improve predictive performance and efficient time to answer cost-effective actions. The author introduced Big Data with its characteristics and Big Data analytics with their types as Descriptive, Predictive and Prescriptive in healthcare industries. Dopamine, a neurotransmitter, generated by brain cells, is responsible to send signals to other brain cells to control muscle activity. The degeneration of dopamine-producing brain cells causes PD. For analysis purposes, voice dataset of PD is collected from UCI machine learning library. By implementing multiple predictive models to disease datasets, multiple accuracies and results of different classifiers are acquired. C4.5, SVM, and ANN give better results than other machine learning algorithms. After comparing the results of these classifiers, best results are chosen for the final decision. This approach helps organizations to analyze their large datasets quickly and efficiently with maximum accuracy.

drksujatha23@gmail.com

Azad et al. [5] explored a predictive model for PD that is based on decision tree algorithm. They introduced PD, a second most common neurodegenerative disease with its symptoms, possible complications, and risk factors associated with it. Various applications of data mining are used for classification purposes that are decision tree, attribute selection measures, ID3 and decision stumps. Their dataset (have 197 instances) is taken from UCI repository and built up from the data of 31 people. For performance analysis, two parameters accuracy and classification error are used. For validation, 10-fold cross-validation technique is used that gives the unbiased outcome. They found that decision tree algorithm performs best and gives the best accuracy and less classification error than other algorithms in their experimental results.

Sriram et al. [6] proposed a method for diagnosis of PD using its voice dataset. This voice dataset is built up from the voice of 31 people among which 23 people are affected by PD. This dataset contains 5875 instances and 26 attributes. In their experiment for statistical analysis, classification, evaluation, visualization, and unsupervised methods Weka V3.4.10 and Orange V2.0b software are used. They achieved the best accuracy 90.2% from Random Forest algorithm.

3 Problem Definition

The huge amount of data known as Big Data is generated everywhere, and this data can be used to perform analysis and make future predictions. According to some research, most of the healthcare data is in the unstructured form that can be stored in the centralized repository to make useful interpretation out of it. To improve the quality of patient care at low cost, this unstructured data can be analyzed further by merging it with structured datasets. The problem is to classify and discover data pattern to predict future disease, so that doctors can detect and diagnose the disease at an early stage.

4 System Architecture

The dataset used in this paper is taken from UCI library [7]. Analysis of these data will provide early diagnosis and detection of disease at reducing cost [4]. The gathered information is in unstructured format, i.e., it is not in a particular kind of format. After that, we convert this mixed type of disease data into the structured form which will ease the process.

For that, a layered Big Data framework is used which has mainly following three components:

Hadoop: A very popular, distributed processing, and storage framework that can handle large and complex unstructured data. Therefore, Hadoop is the best option

drksujatha23@gmail.com

for analyzing unstructured disease dataset. Hadoop has two main components: MapReduce and HDFS for processing and storing a large amount of datasets. Hadoop uses HDFS to store very large data files (in GBs to TBs) that cannot be stored on a single machine. MapReduce is a software programming paradigm used to process a large amount of datasets by distributing the work to various independent nodes.

Predictive Analytics: Is a probabilistic platform for predicting the future. It uses a variety of statistical modeling, data mining techniques, and machine learning techniques. It provides actionable insights based on the data so that organizations can identify a pattern from the data and apply statistical modeling technique and algorithms to find relationships between various datasets [4].

Prediction Models: To classify data various machine learning (ML) algorithms are used. ML (i.e., a branch of Artificial Intelligence (AI) concerned with the study of classification and pattern analysis) algorithm allows us to automatically recognize complex patterns and make intelligent decisions based on data. Some ML algorithms that we used in this paper are as follows:

Support Vector Machine **(SVM)** yields more accurate results when it is used for classifying text. It is successfully employed in text classification and various other sequence processing applications as it is a type of linear classifier.

Artificial Neural Network **(ANN)** is a type of supervised learning models and it is derived from the functionality of human brains, i.e., highly sophisticated analytical technique, capable of modeling extremely complex nonlinear functions [8]. We used a popular ANN algorithm called multilayer perceptron **(MLP)**, i.e., a type of supervised learning model used for prediction and classification problems [9].

5 Proposed Model

We use Waikato Environment for Knowledge Analysis (WEKA) to implement data mining algorithms for preprocessing, classification, clustering, and analysis of results. This environment includes java libraries that implement algorithms and provide the best environment to researchers for classifying datasets.

5.1 Data Collection

The dataset used in this paper is taken from UCI machine learning library [7]. The dataset consists of 195 instances and 24 attributes. This feature set consists of name, Fo(Hz), Fhi(Hz), Flo(Hz), Jitter(%), Jitter(Abs), RAP, PPQ, Jitter:DDP, Shimmer, Shimmer(dB), Shimmer:APQ3, Shimmer:APQ5, APQ, Shimmer:DDA, NHR, HNR, status, RPDE, DFA, spread1, spread2, D2, and PPE. We store these datasets in .CSV format and then convert it into. ARFF format for further analysis. The

drksujatha23@gmail.com

dataset is divided into two classes according to its "status" column which is set to 0 for healthy subjects and 1 for those PD [10].

5.2 Data Preprocessing

The poor data quality in the medical dataset is one of the big challenges that are faced by the knowledge discovery process. This process decreases the number of attributes into a better subset which can increase accuracy, and also it brings a reduction in training time. It is done using Filters and Wrappers. WEKA provides "AttributeSelection" filter to choose an attribute evaluation method. We use "cfsSubsetEval" attribute evaluator and "BestFirstSearch" method which considers the individual predictive ability of each feature to evaluate the worth of an attribute. From that, a new feature data subset is prepared which contains 11 features.

5.3 Data Mining

In this proposed framework, we used different classification techniques for analyzing PD patient's record. To evaluate performance, we apply 10-fold cross-validation technique which splits the original set into training sample to train the model and a test set to evaluate results.

An approach known as "Information Retrieval Metrics" can be used to evaluate experimental results in terms of precision, recall, f-measure, and accuracy with the use of following formulas [11]:

$$\text{Precision} : \text{TP}/(\text{TP}+\text{FP}); \quad \text{Recall} : \text{TP}/(\text{TP}+\text{FP})$$
$$\text{F}-\text{measure} : 2 * \text{Precision} * \text{Recall}/(\text{Precision}+\text{Recall})$$
$$\text{Accuracy} : \text{TP}+\text{TN}/(\text{TP}+\text{TN}+\text{FP}+\text{FN})$$

Here, TP = True Positive; TN = True Negative; FP = False Positive; FN = False Negative [12].

6 Experimental Results

Various techniques are used in the analysis and prediction of PD. Methods that are based on analytics can give an appropriate prediction for a particular disease by grouping people with similar symptoms. In our experiment, obtained accuracies using SMO, KNN [13], Random Forest, AdaBoost.M1 [14], Bagging, MLP, and DT algorithms are 86.67%, 90.76%, 89.23%, 88.20%, 89.23%, and 89.74%,

drksujatha23@gmail.com

Table 1 Performance measured by classifiers

	KNN + AdaBoosta.M1	KNN + Bagging	KNN + MLP
Accuracy	91.28%	90.76	91.28
Classification error	8.717	9.23	8.71
Time taken to build model	0.01	0.02	0.43
Precision	0.873	0.866	0.873
Recall	0.907	0.904	0.907
F-Measure	0.888	0.882	0.888

respectively. In our experiment, we used several ensemble methods that are capable to combine classifiers with their predictions and base estimators. For using more than one classification model, we have to "meta" option under classification tab, and then select "vote" classifier. After that, select the classifier properties and in classifiers tab, we can add multiple classifiers as per our need. KNN provides better accuracy and less execution time than SMO and random forest. So we combined ANN algorithms with KNN algorithm. Table 1 shows performance measure that is reported by our experimental result with disease dataset, after conducting 10-fold cross-validation technique.

Figure 1 shows overall precision, recall, and F-measure rate of combined approach. The accuracy acquired by AdaBoosta.M1 and MLP with KNN are same, i.e., 91.28% which is better when we use these separate algorithms. Our experimental result shows that same preprocessing methods on a different dataset affect similarly the classifiers performance. After analyzing results, it is observed that when we combine two classifiers, accuracy is increased and time taken to build model is reduced. The accuracy acquired by AdaBoosta.M1 and MLP with KNN are same, i.e., 91.28% which is better when we use these separate algorithms. And this accuracy becomes 100% when we use training dataset with selected attributes.

Time taken to build a model with AdaBoosta is 0.01 s, whereas time taken to build model with MLP is 0.43 s. So that AdaBoosta1 with KNN gives best accuracy with time and less classification error.

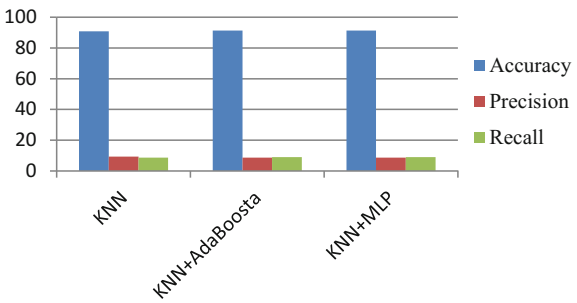


Fig. 1 Comparison of accuracy achieved by ANN algorithms with KNN

drksujatha23@gmail.com

7 Conclusion and Future Scope

Big Data analytics plays a huge role in the healthcare industry, as these data are scattered everywhere, big, and complex in nature. In this paper, we discuss early stage prediction of Parkinson disease for that we presented a methodology of data mining using Weka tool for classifying disease dataset. We use various MLAs for classifying our experimental data that indicate the combined effect of ANN algorithms with KNN which is better as compared when we use other algorithms. The system detects the maximum accuracy of the multi-classifier, and their result predicts the disease at its early stage. We discuss the comparative analysis and calculate the overall performance measures in terms of precision, recall, and f-measure. In future, effective optimization techniques can be used to achieve better accuracy and cost-effective interventions for Parkinson disease. Also, limited data is available that describes the real potential of early PD treatment which requires more research to explore the real impact of early treatment.

References

1. https://en.wikipedia.org/wiki/Parkinson%27s_disease
2. PD: hope through research. https://www.ninds.nih.gov/Disorders/Patient-Caregiver-Education/Hope-Through-Research/Parkinsons-Disease-Hope-Through-Research
3. Saloni et al (2015) Detection of Parkinson disease using clinical voice data mining
4. Shamli N et al (2016) Parkinson's Brain disease prediction using big data analytics
5. Azad C et al (2014) Design and analysis of data mining based prediction model for Parkinson's disease
6. Sriram TVS et al (2013) Intelligent Parkinson disease prediction using machine learning algorithms
7. Shaikh, TA (2014) A prototype of Parkinson's and Primary tumor diseases prediction using data mining techniques
8. Kirubha V et al (2016) Survey on data mining algorithms in disease prediction
9. Salekin A Detection of chronic Kidney disease and selecting important predictive attributes
10. Gaur V et al (2013) A multi-objective optimization of cloud based SLA-Violation prediction and adaptation
11. PD dataset from UCI repository. https://archive.ics.uci.edu/ml/machine-learning-databases/parkinsons/
12. Boukenze B et al (2016) Performance on data mining techniques to predict in healthcare industry: Chronic Kidney failure disease
13. Rana M et al (2015) Breast Cancer diagnosis and recurrence prediction using machine learning techniques
14. Freund Y et al (1996) Experiments with a new boosting algorithm

drksujatha23@gmail.com

Hybrid Technique Based on DBSCAN for Selection of Improved Features for Intrusion Detection System



Akash Saxena, Khushboo Saxena and Jayanti Goyal

Abstract Data mining is the taking out of concealed data from enormous databases (DBs); it is an effective innovation with unusual probable to enable associations to deliberate on the mainly imperative data in their data warehouses. IDS are the chief issue of the security which is helpful in everyday life to avoid the data from the attackers. Data mining includes numerous methods for the detection of intrusion which involves the detection of all harmful activities. In our proposed work, we initially apply KDD cup'99 dataset which is most broadly used method for detecting intrusion. DBSCAN is the most utilized method which is used to eliminate noise from the data. Then, we generate the most meaning inputs by analyzing and processing whole data which is done by the selection of feature method. K-means clustering performs grouping of data which is followed by SMO classifier. So we proposed a hybrid structure which improves the taken as a whole accuracy. MATLAB and WEKA tools are used to execute the whole process.

Keywords Data mining · Knowledge discovery database · Machine learning SVM · K-means

A. Saxena
Department of Computer Science & Engineering, Compucom Institute of Information Technology and Management, Jaipur, India
e-mail: akash27saxena@gmail.com

K. Saxena (✉)
Department of Computer Science & Engineering, Oriental Institute of Science & Technology, Bhopal, India
e-mail: kskhushboosaxena26@gmail.com

J. Goyal
Kanoria PG Mahila Mahavidyalaya, Jaipur, India
e-mail: goyal.jayanti@gmail.com

© Springer Nature Singapore Pte Ltd. 2019

V. S. Rathore et al. (eds.), *Emerging Trends in Expert Applications and Security*,
Advances in Intelligent Systems and Computing 841,
https://doi.org/10.1007/978-981-13-2285-3_43

drksujatha23@gmail.com

1 Introduction

The information is accessible in the distinctive measures with the goal that the best possible interchange of data to be made. To analyze this information as well as take a decent choice and carry on the information, as and when the client will need, the information ought to be recovered from the DB and resolved the healthier conclusion. There is an enormous volume of data; anyway, we barely prepared to hand them over to supportive information and learning for authoritative essential making in business. To take the complete preferred standpoint of information, the data recovery is essentially enough; it requires a gadget for the customized diagram of data, extraction of the core of dataset away, and the revelation of examples in raw data. With the immense computation of dataset away in archives, DBs, and differing stores, it is always essential, to create the capable gadget for examination and interpretation of such data and for the extraction of captivating determining that might assist in elementary leadership [1].

2 Data Mining Techniques

2.1 Knowledge Discovery in Databases

Data mining is the course of breaking down data from exchange points of view and compacting it into significant data. Actually, data mining is the course toward discovering associations or examples among numerous fields in significant social databases.

The figure "Fig. 1" roars demonstrate the diverse strides for separating valuable data from volume data.

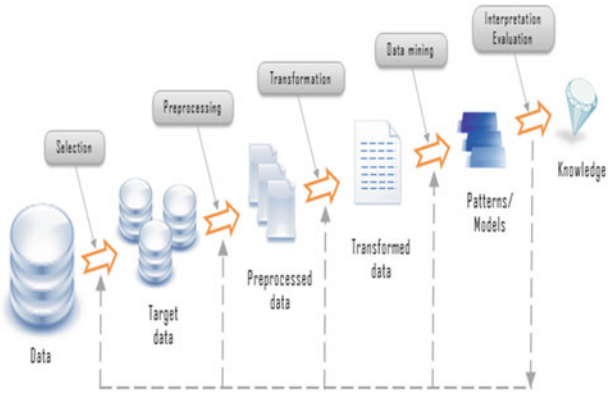


Fig. 1 Overview of the steps constituting the KDD process

drksujatha23@gmail.com

2.2 Data Mining Techniques

In this segment, we exhibit the notable data mining systems that have generally utilized as a part of intrusion detection.

2.3 Machine Learning

Machine learning is one of the examinations of synthetic knowledge fields, is the intelligent train stressed over the advancement, investigation, and usage of robotized strategies that enable a machine to develop through a procedure of learning, and to complete errands that are troublesome or difficult to fill by classic algorithmic.

(a) **Clustering techniques or algorithms**: The rule for checking the likeness is depending on calculation to other and from grouping model to other. The hierarchical calculation in light of division arranges in this way the availability models.

(b) **Support vector machine (SVM)**: SVM has been associated with different fields, for instance, bioinformatics, information recovery, PC vision, etc. According to the information, the execution of SVM is tantamount, or even unrivaled, to the neural network or a Gaussian mixture model.

(c) **Genetic estimation (GA)**: Also, it is a field of computerized reasoning. GA has been associated in various fields with promising outcomes to optimum results. Genetic algorithms depend on natural marvels. Genetic has uncovered the presence of a few tasks inside a living being offering ascend to genetic mixing.

- **Selection**: To figure out which people will probably get the best outcomes, a choice is made.
- **Crossover**: During this activity, two chromosomes trade bits of their chains, to give new chromosomes. These hybrids might be single or various.
- **Mutation**: Randomly, a quality can inside a chromosome being substituted for another. A vague course for hybrids; here, we describe a change rate while changing people that is generally in the region of 0.001 and 0.01.

(d) **Fuzzy logic**: It is an expansion of established rationale to inexact thinking created by Lotfi Zadeh in 1965, in light of an arrangement of non-computerized set hypothesis and standards. Numerous different creators utilize fuzzy data mining way to deal with extricating designs which introduce ordinary or malevolent conduct for intrusion detection frameworks (Fig. 2).

drksujatha23@gmail.com

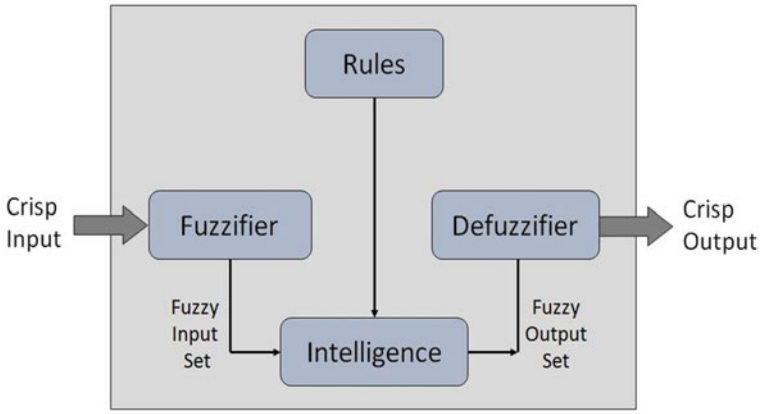


Fig. 2 Fuzzy logic system

3 Data Mining Application

Distinctive field adapted data mining progresses by virtue of brisk access to information and essential information from a great deal of information. Data mining application domain consolidates showcasing, media transmission, misrepresentation recognition, fund, and training territory, therapeutic and whatnot. A bit of the rule applications recorded underneath:

(a) **Data Mining in Education Sector**: We are applying information mining in training area than new rising field called "Instruction Data Mining" [2].
(b) **DM in Banking and Finance**: [3] in the managing an account field, information mining is used to anticipate Visa misrepresentation, to evaluate possibility, and to research the example and profitability.
(c) **Data Mining in Market Basket Analysis**: The stores can utilize this data by putting these items in the closeness of each other and making them more unmistakable and open for clients at the season of shopping [4].
(d) **Data Mining in Earthquake Prediction**: Earthquake is the abrupt advancement of the Earth's outside layer caused by the surprising entry of stress gathered next to a geologic blunder in within [5].
(e) **Data Mining in Agriculture**: Data mining than creating in agribusiness field for altering yield investigation with respect to four parameters particularly year, precipitation, age, and zone of sowing. Yield forecast is a basic agrarian concern that outstanding parts to be handled in light of the available details [6].
(f) **Data Mining in Cloud Computing**: The utilization of DM methodology through Cloud processing will empower the customers to recoup huge information from in every practical sense coordinated data distribution center that decreases the costs of establishment and capacity [7].

drksujatha23@gmail.com

4 Goals of Data Mining

4.1 Prediction

Prediction chooses the relationship involving autonomous factors and association among dependent and independent features.

4.2 Identification

Data patterns are required to make out the existence of the item, an event, or several patterns those are of customer behavior. The identified area is as authentication is the layout of classification.

4.3 Classification

DM can help to divide the data so that various classes can be recognized based on the parameters grouping to search a clever say that to show data.

4.4 Optimization

DM can enhance the utilization of resources those are incomplete, such as time, space, cash, or materials and to enlarge output those factors.

5 Advantages of DM

DM approaches suppose an essential element in the dissimilar domain. For the categorization of issues of security, a lot of details ought to be analyzed including the data for verifiable. It is bothersome for people to determine an example of such a huge quantity of data. DM, in several cases, come out to be suitable to crush this difficulty and can be utilized to choose those models [8].

drksujatha23@gmail.com

6 Literature Survey

Tyagi and Jawdekar [9] think about the impact of web-based business on business sectors where found organizations confront hopeful from web built-up incoming with the focused alternative. They comprised and developed a recommender system to dismember the achievement of contenders on the items. The marked down things are provided to the clients with an end goal to grow the customer's interest. With more than 200 million clients online to the www, the advanced business now account owed for a creating level of world trade [9].

Groeschel [10] showed that IDS screen system or host bundles endeavoring to recognize poisonous activities on a structure. Peculiarity discovery frameworks have accomplishment in revealing new attacks, usually implied as "zero"-day assaults, and however, comprise high false positive rates [10].

Gaied et al. [11] exhibited that computer security is a long way from being ensured because of the adaptability of PC arranges, the steady development of dangers, and the nearness of boisterous data. The central objective of the present work is first to misuse information mining frameworks. The subsequent purpose of this paper is the knowledge limit of neural systems, and the other one is the fuzzy rules suspecting that recognizes data wave characters [11].

Ait Tchakoucht et al. [12] displayed that IDS are frequently used to accumulate and investigate network traffic to empower heads to design and oversee assaults. In behavioral approach, these identification frameworks manage the entire framework to distinguish anomalies in the wake of working up the system common profile including all clients [12].

Alseiari et al. [13] displayed that, as AMI portions are related through work organizes in a dispersed instrument, latest vulnerabilities will be mishandled by matrix's assailants who deliberately intrude with system's correspondence structure and take client information [13].

Desale et al. [14] introduced that the recent developing growth of data made such a large number of difficulties in data mining. Data mining is the way of separating legitimate, already known and extensive datasets for the future basic leadership [14].

Elekar et al. [15] presented that, as the Internet continues to influence our day-to-day activities like e-Commerce, e-Governance, e-Education, etc., the danger from hackers has additionally expanded, because of which numerous researchers think IDS as the crucial line of defense. Nonetheless, numerous economically accessible IDSs are predominantly signature-based that are intended to detect known attacks [15].

Leu et al. [16] displayed that, currently, most PC frameworks utilize client IDs and passwords as the login cases to approve customers. Insider attackers, the legitimate clients of a system who attack the structure inside, are difficult to distinguish while nearly all interruption identification frameworks and firewalls recognize and segregate pernicious practices propelled from the outside universe of the framework as it were [16].

drksujatha23@gmail.com

Ng et al. [17] exhibited that, in our present society, the danger of digital intrusion is progressively very highly dangerous. With the ascent of utilization in PCs, an illicit movement has additionally moved from physical interruption into the digital interruption. Current strategies utilized for these frameworks incorporate utilizing inconsistency recognition or a signature database [17].

Subaira and Anitha [18] introduced that, regardless of developing data framework broadly, security has stayed one hard-hitting territory for PCs and moreover organizes. In information protection, IDS is used to ensure the information classification, respectability, and framework accessibility from different kinds of attacks [18].

Hasija and Chaurasia et al. [19] to consult clients with a race of using a web content recommender framework which delivers customized recommendation timelily relied upon their very own imitation practices and propensities. And to start with, data purifying is achieved and after that semantically upgraded web utilization logs are prepared. At that point, web get to conduct is perceived by asset and occasional traits. Fuzzy C-Means (FCM) clustering calculation has been connected and relies upon certainty and support esteems, the rambling example is accomplished intersection a specific edge [19].

Modi and Narvekar et al. [20] in this paper, an engineering which incorporates item information with client get to log information and after that creates a gathering of the proposal for which demanding client is introduced. The usage has recorded engaging terms of exactness, review, and F1 estimations [20].

7 Discussion and Result

KDD'99 cup datasets are used for intrusion detection metrics. Each TCP connection is represented by 41 features. It contains 4,940,000 association records. Every TCP association was named as "normal" or "attack "with a particular attack write"; the length of every association record is 100 bytes. Simulated attack composes fall in one of the accompanying four classifications.

7.1 Denial-of-Service Attack (DOS)

In this sort of *attack*, the figuring or memory asset turns out to be excessively occupied or too full, making it impossible to deal with true blue demands, or denies honest to goodness clients access to a machine.

drksujatha23@gmail.com

7.2 User to Root Attack (U2R)

In this, *attacker* get to an ordinary customer account on the structure and can abuse some shortcoming to get root access to the system.

7.3 Remote to Local Attack (R2L)

It happens when an *attacker* who can send parcels to a machine over a system yet who does not have a record on that machine abuses several feeblenesses to increment adjacent access as a customer of that machine.

7.4 PROBE Attack

It is an endeavor to assemble data about a *network* of PCs for the obvious reason for bypassing its security controls. *Attacks* could be grouped in light of the *combination* of these 41 *features*.

In the proposed method, there is necessitate to choose only computable feature so we utilize feature selection, and it as a case of a study recently has become the significant phase of enhancement of an IDS as it affects the performance of a classifier. Methods of feature selection can effectively recognize a features subset within a dataset and diminish the quantity of fields, in order to decrease the time for computation process. The KDDCUP 99 dataset includes 41 attributes but we use the only preferred attribute which we attain from the selection of feature. Presently,

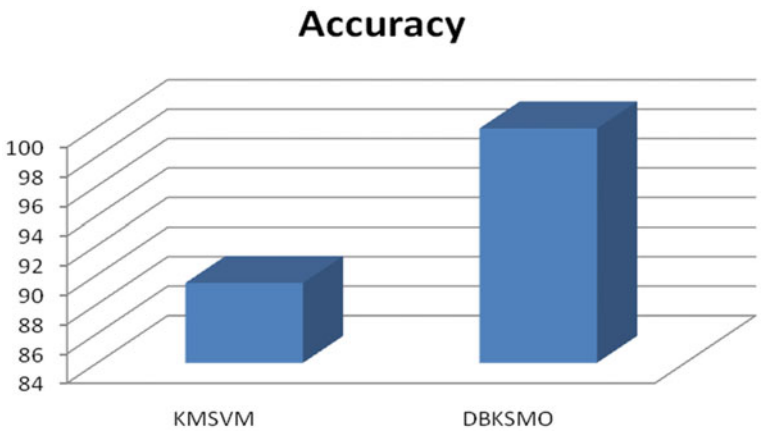


Fig. 3 DoS attack

drksujatha23@gmail.com

in initial step, we utilize DBSCAN to eliminate noise from dataset, so we obtain noise-free data for the further process. In next step, K-mean is functional on training data and as outcome clusters are created for DOS, PROB, R2L, and U2R attack. In the last step, SMO is utilized as a classifier to check whether interruption happened or not. Following figures demonstrate that the exactness appears concerning for all attack categorization classes attained from mixture and hybrid system DBKSMO. For this, the dataset utilized does not enclose each one of the 41 traits; somewhat, it surrounds presently selected attribute set. KMSVM stands for K-means with support vector machine which is used for the detection of the intrusion and it is compared with DBKSMO. Here, we can see that DBKSMO achieves enhanced as the contrast to other algorithms KMSVM (Figs. 3, 4, 5, and 6).

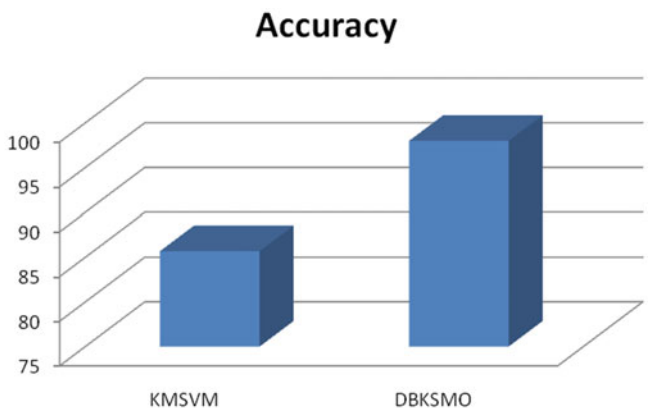


Fig. 4 PROB attack

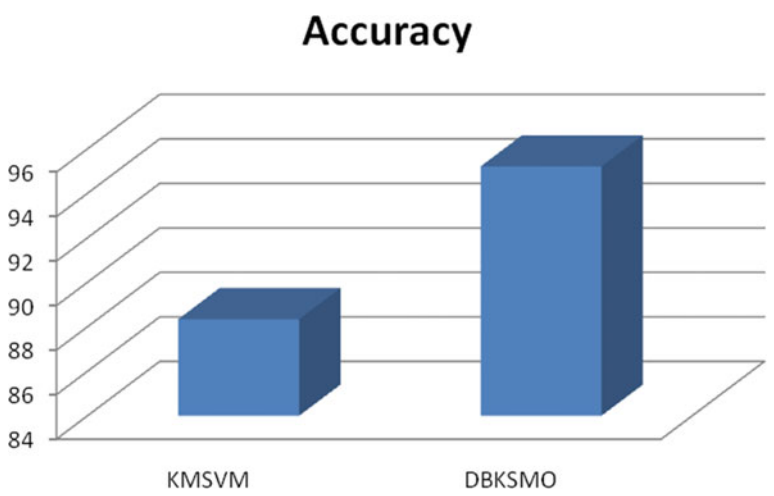


Fig. 5 U2R attack

drksujatha23@gmail.com

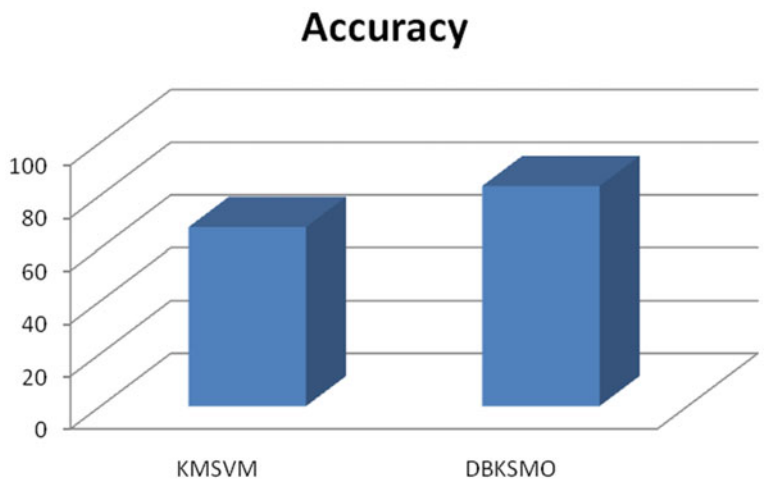


Fig. 6 R2L attack

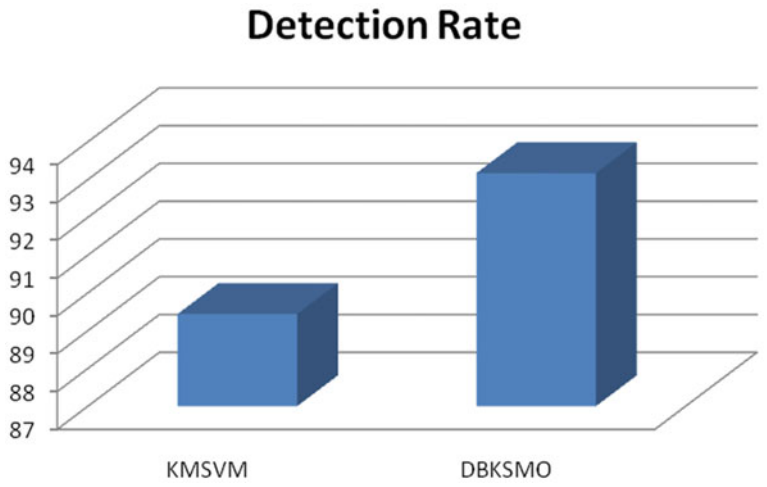


Fig. 7 DoS attack

Following figures show that the rate of detection appears to fruition for the entire attack characterization classes while we have utilized the testing dataset including just picked properties. Besides these lines, the detection rate of DBKSMO is much improved than as contrast with KMSVM calculation (Figs. 7, 8, 9, and 10).

drksujatha23@gmail.com

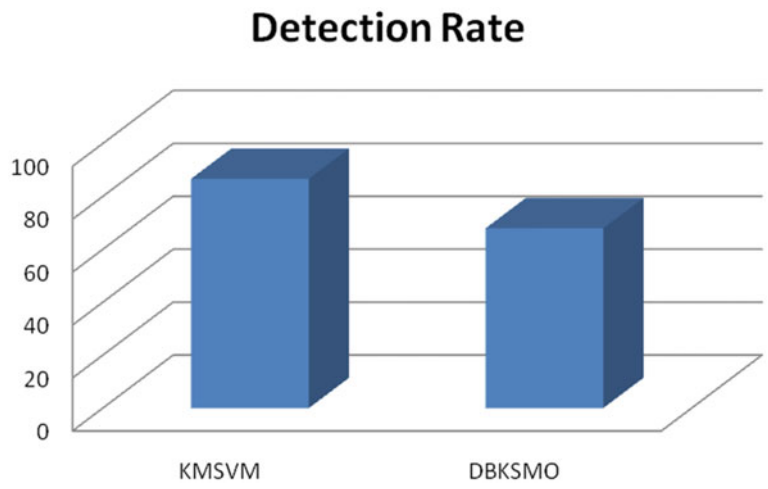


Fig. 8 PROB attack

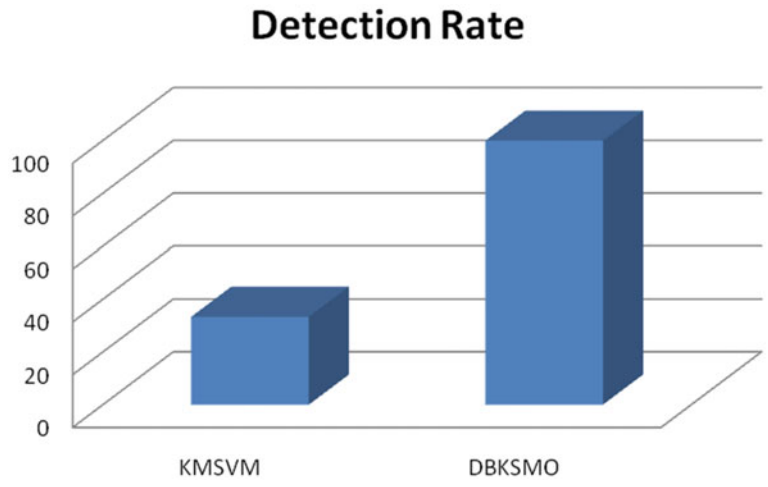


Fig. 9 U2R attack

drksujatha23@gmail.com

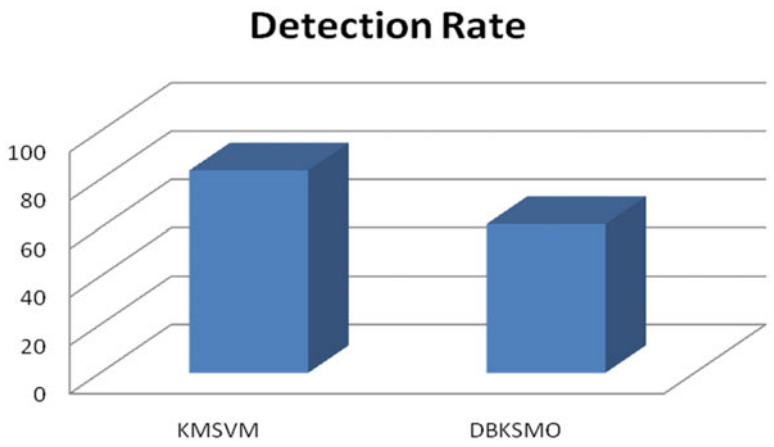


Fig. 10 R2L attack

8 Conclusion

IDS monitors network or host parcels endeavoring to recognize vindictive actions on a structure. Anomaly detection frameworks comprise accomplishment in revealing latest attacks, usually insinuated as "zero"-day attacks, yet include positive rates of high false. The exploration in data stream mining and IDS expanded high fascination appropriate to the centrality of system's well-being evaluation. Calculations, systems, and structures that tackle challenges of security have been prepared over the earlier days. Security is the critical component for data in the present instance which makes IDS become critical for the detection of various actions. We executed DBScan, clustering, selection of features, and SMO for the detection of intrusion. In addition, we utilized K-means which clusters the data into four groupings to improve the accuracy of the entire technique. This method provides enhanced result than the existing work which can be exposed by the graphs. From the proposed work, the accuracy and detection rate of the method have enhanced to a superior scope.

References

1. Padhy N, Mishra P, Panigrahi R (2012) The survey of data mining applications and feature scope, vol 2, No 3, June 2012. https://doi.org/10.5121/ijcseit.2012.2303
2. Umamaheswari K, Niraimathi S (2013) A study on student data analysis using data mining techniques. Int J Adv Res Comput Sci Soft Eng 3(8)
3. Industry Application of data mining. http://www.pearsonhighered.com/samplechapter/0130862711.pdf
4. Olson DL, Delen D (2008) Advanced data mining techniques. Springer

drksujatha23@gmail.com

5. Otari GV, Kulkarni RV (2012) A review of application of data mining in earthquake prediction GV Otari et al/(IJCSIT). Int J Comput Sci Inf Technol 3(2):3570–3574
6. Ramesh D, Vardhan BV (2013) Data mining techniques and applications to agricultural yield data. Int J Adv Res Comput Commun Eng 2(9)
7. Petre R-Ş (2012) Data mining in cloud computing. Database Syst J III(3)
8. Jaiganesh V, Mangayarkarasi S, Sumathi P (2013) Intrusion detection systems: a survey and analysis of classification techniques. Int J Adv Res Comput Commun Eng 2(4). ISSN (Print): 2319–5940
9. Tyagi R, Jawdekar A (2016) An advanced recommendation system for e-commerce users, 978-1-5090-0669-4/16/$31.00 © 2016 IEEE
10. Goeschel K (2016) Reducing false positives in intrusion detection systems using data-mining techniques utilizing support vector machines, Decision Trees, And Naïve Bayes For Off-Line Analysis, 978-1-5090-2246-5/16/$31.00 © 2016 IEEE
11. Gaied I, Jemili F, Korbaa O (2015) Intrusion detection based on neuro-fuzzy classification, 978-1-5090-0478-2/15/$31.00 © 2015 IEEE
12. Ait Tchakoucht T, Ezziyy Ani M, Jbilou M, Salaun M (2015) Behavioral approach for intrusion detection, 978-1-5090-0478-2/15/$31.00 © 2015 IEEE
13. Alseiari FAA, Aung Z (2015) Real-Time anomaly-based distributed intrusion detection systems for advanced metering infrastructure utilizing stream data mining, 978-1-4673-8734-7/15/$31.00 ©2015 IEEE
14. Desale KS, Kumathekar CN, Chavan AP (2015) Efficient Intrusion Detection System using Stream Data Mining Classification Technique, https://doi.org/10.1109/iccubea.2015.98, 978-1-4799-6892-3/15 $31.00 © 2015 IEEE
15. Elekar KS (2015) Combination of data mining techniques for intrusion detection system. In: IEEE International Conference on Computer, Communication and Control (IC4-2015)
16. Leu Fang-Yie, Tsai Kun-Lin, Hsiao Yi-Ting, Yang Chao-Tung (2015) An internal intrusion detection and protection system by using data mining and forensic techniques. Digit Object Identifier. https://doi.org/10.1109/JSYST.2015.2418434IEEE
17. Ng J, Joshi D, Banik SM (2015) Applying data mining techniques to intrusion detection, https://doi.org/10.1109/itng.2015.146, 978-1-4799-8828-0/15$31.00 © 2015 IEEE
18. Subaira AS, Anitha P (2014) Efficient classification mechanism for network intrusion detection system based on data mining techniques: a survey, 978-1-4799-3837-7/14/$31.00 © 2014 IEEE
19. Hasija H, Chaurasia D (2015) Recommender system with web usage mining based on fuzzy C means and neural networks. In: 2015 1st International Conference on Next Generation Computing Technologies (NGCT-2015) Dehradun, India, 4–5 Sept 2015. IEEE
20. Modi HY, Narvekar M (2015) Enhancement of online web recommendation system using a hybrid clustering and pattern matching approach, 978-1-4799-7263-0/15/$31.00 © 2015 IEEE

drksujatha23@gmail.com

A Study on Performance Evaluation of Cryptographic Algorithm



Mohammed Firdos Alam Sheikh, Sanjay Gaur, Hiral Desai and S. K. Sharma

Abstract Nowadays, many security algorithms are used for security purpose. Encryption provides a solution in security system. In recent years, security of network had important issue. My research work shows the performance of the encryption methods like AES, DES, and RSA algorithms. By experiment result, it is concluded that RSA takes longest time for encryption and AES algorithm take shortest encryption time. We also conclude that decryption of AES also is excellent in comparison with other algorithms. There is no single method that will provide all the services specified. In this research, we identify that an important mechanism which will help different types of integrity is called cryptographic technique.

Keywords Encryption · DES · AES · RSA · Cryptography

1 Introduction

The purpose of cryptography is the security of the data from unauthorized access. Cryptography is a technique of protecting data by encrypting it into an unreadable form. This information can only be accessed by those who have the secret key that can decrypt the information into plaintext (original form). Cryptography is an

M. F. A. Sheikh (✉)
PAHER, Udaipur, India
e-mail: firdos.sheikh@gmail.com

S. Gaur
Jaipur Engineering College and Research Centre, Jaipur, India
e-mail: sanjaygaur.cse@jecrc.ac.in

H. Desai
Pacific School of Engineering, Surat, India
e-mail: hiral8.desai@yahoo.com

S. K. Sharma
MITRC, Alwar, India
e-mail: Skpacific323@gmail.com

© Springer Nature Singapore Pte Ltd. 2019

V. S. Rathore et al. (eds.), *Emerging Trends in Expert Applications and Security*,
Advances in Intelligent Systems and Computing 841,
https://doi.org/10.1007/978-981-13-2285-3_44

drksujatha23@gmail.com

important part of securing private data from an unauthorized world. Symmetric encryption techniques are faster than asymmetric techniques as they require less computational time. For secured interface, over wild network information can be secured by the encryption method. Encryption technique translates that data using an encryption algorithm using the key in cipher (unreadable) form. In symmetric key encryption, only one key is used to encrypt and decrypt data. Information which is easily read and understand by user is knowing as plaintext. The method of converting the plaintext into non-understandable form by any algorithm is knowing by encryption. Using encryption algorithm, plaintext changes the message into non-understandable form that is known as ciphertext. The technique of deciphering the ciphertext to its readable form is knowing as decryption [1].

2 Cryptography

Cryptographic algorithms are categorized as symmetric algorithms (secret key) and asymmetric algorithms (public key). Same key is used in symmetric key encryption by both sender and receiver. In asymmetric key encryption, two distinct keys are used: one is public key which is used for encryption and other key is private key which is used for decryption. The examples of asymmetric encryption algorithms are RSA and ECC. There are five ingredients of cryptography:

(i) Plain text

Plain text is the readable data which is given to the algorithm for input.
b. Encryption Algorithm
Plain text can be converted into ciphertext.

(ii) Secret key

It applies the input to the algorithm, and its value is not dependent on the plaintext.
Encryption is a technology for preserving important data. In this method, public and private key encryptions are used to prevent the confidential data [2].

3 Symmetric Encryption Schemes

In block ciphers, put an input the key and a block, it uses the same size as the key. The first block is developed by the initialization vector, which provides some addition of randomizing to the encryption.

drksujatha23@gmail.com

3.1 DES Algorithm

It is based on Data Encryption Standard (DES). We have two inputs for encryption function, the key and PT is to be encrypted. A plain text should be 64 bits in length and 56 bits of key. First, the plain text of 64 bits passes by an initial permutation which reforms the bits. This is having 16 rounds of same function, which associate permutation and substitution functions. After complete of 16 rounds, the output is interchanged at 32 bits position.

DES algorithm consists of the following steps.

3.2 Encryption

DES allows 64-bit large plaintext as an input and 56-bit length of key and then generates output of 64-bit block.

The block of plaintext had shifted the bits around.

From the key, there are 8 parity bits removed to its key permutation.

The key is partitioned into two 28–28 bits.

Half of each key is shuffle by 1 or 2 bits.

For minimizing the key from 56 bits to 48 bits, the halves are rejoined to a compressed permutation. This shorten keys is used to encrypt that round of plaintext block.

The rotated key section from step two is used in the next round.

The block is partitioned into 32–32 bits of two fractions.

One halve passes through an expansion permutation to increase the block size by 48 bits.

Result from step 6 is XOR'ed with the 48-bit shorten key from step 3.

Result from step 7 is sent to substitution box, which minimizes from 48-bits of block to 32 bits of block (Fig. 1).

The result of step 8 is passed through permutation box to permute the bits.

The result from the permutation box is XOR'ed with other block.

The data of two divisions are interchanged and it will be used as the input of the next rounds.

drksujatha23@gmail.com

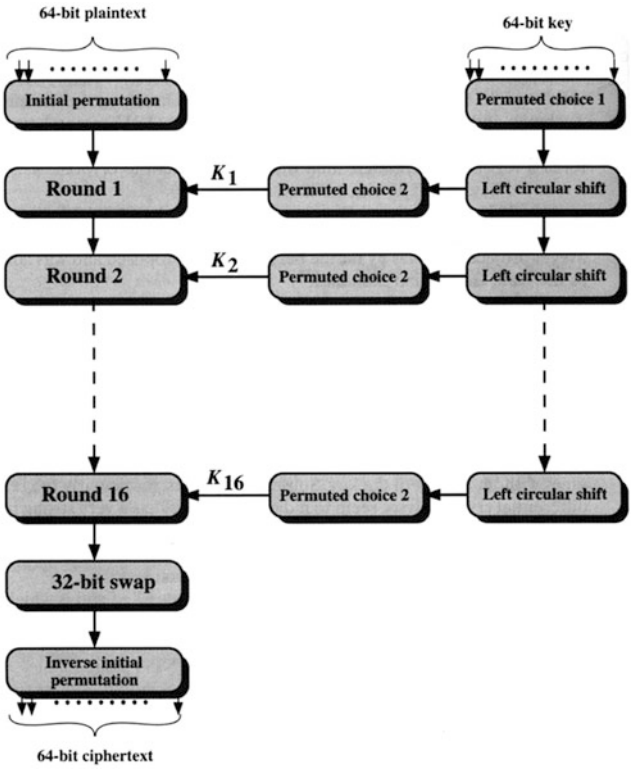


Fig. 1 DES algorithm

3.3 AES

AES algorithm is used for security as well as speed. It encrypts the blocks of data of 128 bits in 10, 12, and 14 rounds which depends on the key size. This could be executed on different platforms. It is verified for various security applications.

4 Present Work

We take four text files in which sizes are different to be done for trials; a differentiation of three algorithms are AES, DES, and RSA. The experimental result for encryption algorithm is shown in Table 1 and Figs. 2, 3.

drksujatha23@gmail.com

Table 1 Comparison of evaluation time of DES, AES, and RSA

S.nO	Algorithm	Packet size (KB)	Encryption time (S)	Decryption time (s)
1	AES	163	1.7	1
	DES		3.1	1.1
	RSA		7.7	5.2
2	AES	210	1.8	1.5
	DES		2.1	1.32
	RSA		9.1	6.3
3	AES	330	1.9	1.6
	DES		3.1	1.37
	RSA		8.2	5.3
4	AES	905	2.0	1.8
	DES		4.1	1.25
	RSA		8.5	5.3

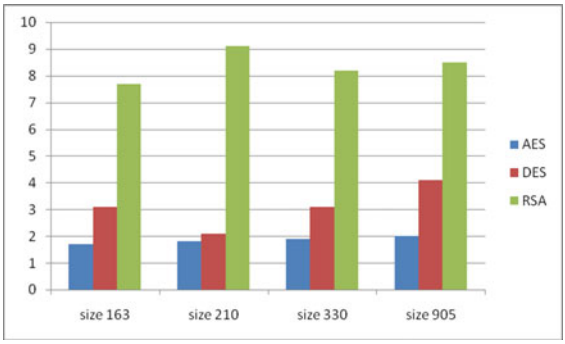


Fig. 2 Comparison of encryption time among AES, DES, and RSA

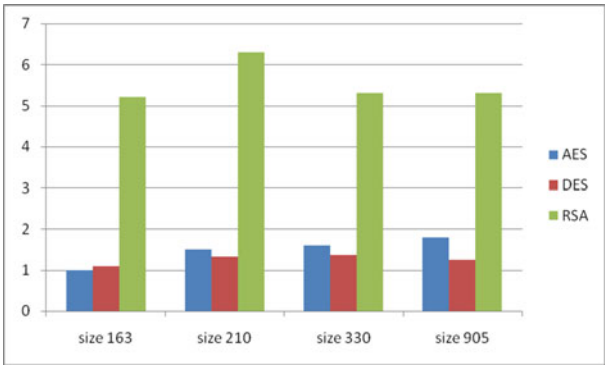


Fig. 3 Analogy of decryption time between AES, DES, and RSA

drksujatha23@gmail.com

5 Conclusion

Our work shows the performance of the encryption algorithms like AES, DES, and RSA. By experiment result, it is concluded that RSA takes longest time for encryption and AES algorithm takes shortest encryption time. We also conclude that decryption of AES algorithm is excellent in comparison of other algorithm. From this, we have concluded that AES algorithm is far better than DES and RSA.

6 Future Scope

This paper gives the valuable knowledge in understanding the cryptographic algorithms of its basic work and which algorithm is performed well than others. In this paper, performance evaluation is done on only AES, DES, and RSA algorithms but this can be furthermore analyzed in future by taking other parameters or algorithms. And it will be interesting to study how performance can be affected by taking different parameters.

References

1. Schneider B Applied cryptography: protocols, algorithms, and source code in C, 2nd ed. Wiley
2. Finnigan P (2002) SQL injection and oracle—parts 1 & 2. Technical Report, Security Focus, Nov 2002. http://securityfocus.com/infocus/1644

drksujatha23@gmail.com

Optimal Ant and Join Cardinality for Distributed Query Optimization Using Ant Colony Optimization Algorithm



Preeti Tiwari and Swati V. Chande

Abstract Query Optimization in Distributed Database Management System (DDBMS) involving large number of relations with multiple joins has always been an attractive area of research. Ants are the social agents in Ant Colony Optimization Algorithm that are responsible for generating optimized solutions to the problem under study. The appropriate numbers of ants needed to generate optimal solutions in terms of both join cardinality, response time is continuously under consideration by researchers as small number of ants leads to premature convergence, and large number of ants leads to high exploration causing slower convergence. This paper attempts to estimate minimum number of ants needed to optimize distributed queries with varied number of joins. This estimation is coined as Ant Ratio, which evaluates the requirement of x number of ants for optimizing distributed query with y number of joins.

1 Introduction

Distributed Database (DDB) is described as a single logical database stored on multiple interconnected computers that may be geographically separated [5]. The Distributed Database Management System (DDBMS) offers higher data availability, reliability, and security to data and enhances the performance and efficiency of the system by implementing Transparency and Local Autonomy [3]. A query in Distributed Environment includes processing, manipulating, and retrieving data from large number of relations stored at multiple locations with the help of Local and Global Database Manager. When the query enters the DDBMS environment, it is parsed, validated, and decomposed into algebraic form based on global conceptual schema. The Localization Manager restructures the query onto local con-

P. Tiwari (✉) · S. V. Chande
International School of Informatics and Management, Jaipur, Rajasthan, India
e-mail: preeti.tiwari@icfia.org

S. V. Chande
e-mail: swatichande@icfia.org

© Springer Nature Singapore Pte Ltd. 2019

V. S. Rathore et al. (eds.), *Emerging Trends in Expert Applications and Security*,
Advances in Intelligent Systems and Computing 841,
https://doi.org/10.1007/978-981-13-2285-3_45

drksujatha23@gmail.com

ceptual schema and creates fragmented query. By permuting the order of operations, many equivalent query execution plans are generated with some *better* than others in terms of execution cost and execution time [12, 16]. These equivalent query execution plans are received by Global Distributed Query Optimizer (GDQO) that evaluates all permutated query plans based on data transmission statistics between sites [15, 19, 21]. The major objective of GDQO is to find the "best" query execution strategy among the candidate solutions to ensure quick, effective, accurate, and reliable results with minimum utilization of the system resources.

Amongst the various evolutionary algorithms proposed and implemented so far by eminent researchers, Ant Colony Optimization Algorithm (ACO) has gained considerable attention as GDQO to solve query optimization problem of RDBMS and DDBMS [1, 2, 11, 13, 14, 18]. This is because of the constructive approach of the algorithm to evaluate each option dynamically at the time of execution and create better plans rather than to first create a plan and then evaluate and improve. Ants carry out the solution construction phenomenon of ACO. To estimate the minimum number of ants needed to optimize a distributed query where the number of relations exhibiting join operations changes with every input query is still a task under consideration. The major objective of this paper is to determine the minimum number of ants needed to optimize a distributed query with varied number of joins. Section 2 introduces the significance of ants in ACO, Sect. 3 elaborates the experimental setup, Sect. 4 includes results and analysis, and Sect. 5 concludes the findings.

2 Significance of Ants in Ant Colony Optimization Algorithm

Ant Colony Optimization was proposed by three Italian scholars, Dorigo et al. [4] as a Combinatorial Optimization Algorithm grounded on foraging behavior of real ants to discover the shortest path between their nest and food source. It uses probabilistic technique and constructive approach for solving large computational problems [8]. ACO exhibits strong characteristics like optimistic search techniques, parallel processing with greedy approach to generate global optimized results. The algorithm iterates over three major segments. In the *ConstructAntSolution* segment, set of m simulated ants construct solutions from elements of a finite set of available solution components. The *ApplyLocalSearch* segment improves the solutions locally by applying probabilistic techniques before updating the pheromones and the *UpdatePheromone* stage increases the pheromone values of the promising solutions, and decreases the pheromone values of the bad solutions [7]. To implement ACO as an optimizer in a DDBMS environment, the query is represented as a connected construction graph $G = (N, E)$ where N denotes nodes

drksujatha23@gmail.com

representing the relations involved in query and *E* denotes the edges representing the join between these relations because ACO works best when problem is represented in the form of graphs. The length of the edges connecting the vertices is equivalent to the cardinality of the relations and each vertex is associated with the pheromone values and heuristics values. The time complexity of ACO is $O(N_c * n^2 * m)$ and Space Complexity $S(n) = O(n^2) + O(n*m)$ [10] where *N* represents the total number of iterations to generate the solution, n represents the problem size, and *m* represents the total number of ants used to generate the optimal solution.

Ants are social agents of ACO that live in a discrete world and are responsible for searching, constructing, and improving solution sets to generate *Optimized Solutions* [9]. The ants initially wander randomly in all directions to find food and return to their colonies. During this search, they deposit a chemical on their path called Pheromone to attract other ants on the same path. If other ants find similar path, they stop traveling randomly and follow the trail. The longer the time taken by ant to travel down the path to find food, the higher is the evaporation of Pheromone and longer is the path. In comparison to this, shorter path is marched more frequently by ants thus having higher density of Pheromones. The result is that all ants follow one single path. The ACO impersonates this behavior of ants with simulated ants walking around the graph, searching for optimal path for the problem to solve [6, 7].

To optimize queries in distributed database, the artificial ants incrementally build solutions by moving on the construction graph biased by a Pheromone model where the parameter values on the graph are modified by ants at runtime Gambardella et al. [10]. The ants keep track of the vertices it has visited and apply probabilistic rule to choose the next vertex. This depends on the heuristic value, shared memory containing experience gathered by the ants in the previous iteration and pheromone concentration value. Higher values show higher probability of selection [7, 8]. They indirectly interact or communicate with each other through pheromone concentration where one modifies the environment by increasing pheromone concentration and the others respond to it by adapting and following the higher concentration paths. These ants deposit pheromone in a problem independent way. The total number of ants performing the search operation to generate optimal solutions plays a significant role. If the number of ants is extremely less as compared to the problem size, incomplete search takes place and the algorithm falls into local optimum leading to premature convergence. If large number of ants is involved in generating the optimal solutions, the randomness in the exploration increases leading to enhanced response time. This paper focuses on the estimation of optimal number of ants needed for optimizing distributed queries. The main objective of this experiment is to find out a stabilization point where adding extra ants to the problem will not affect the quality of the solution sets.

drksujatha23@gmail.com

3 Experimental Setup

The simulation system is designed to optimize extremely large join queries in distributed database with minimum number of ants. For each relation, the algorithm generates an arbitrary set of tuples at runtime that perform the join operations. Based on the number of joins involved in the fragmented query, the algorithm creates a search space that consists of multiple ways of executing the join operation. These Join Orders differ from each other in terms of Data Statistics and Execution Time. The ants search, construct, and retrieve the optimal Join Order that involves minimum number of tuples in minimum response time. The total number of ants performing this search operation plays a significant role. Table 1 enlists the other parameters of ACO used for effective implementation along with their defined values. The values of these parameters affect the performance of the algorithm as they affect the global convergence and solution efficiency [17, 20].

The algorithm accepts number of joins as an input that varies from 25 to 100 with an interval of 5. For each join number, ACO is executed with varied number of ants ranging from 1 to 30 (with an interval of 1). For every query executing join operation involving x number of joins, the algorithm executes for different number of ants starting from a single ant solving the query optimization problem to 30 ants solving the same query optimization problem. The algorithm executes and evaluates the Optimal Cardinality of the Join Operation along with the time taken for evaluation and generation of optimized Join Order. In order to test the effectiveness and performance, a computer with Intel® Core ™ i5-4200U CPU @ 1.60 GHz 2.30 GHz and 6 GB RAM specifications is considered for experiments.

Table 1 ACO discrete parameter value

Symbol	Parameter name	Description	Value
(α)	Pheromone relativity importance	Amount of information accumulated (Stochastic Factors) in the ants movement on their trial paths	1
(β)	Heuristic value	Information needed by ants for their movement on trial paths	5
(ρ)	Pheromone evaporation coefficient	Residue of pheromone left by an ant on the edge that evaporates with time	0.1
(q)	Pheromone deposition coefficient	Amount of pheromone to be deposited on the trial paths traversed by ants	2
TC	Termination criterion	Maximum iterations	100

drksujatha23@gmail.com

4 Results and Analysis

Table 2 displays the Solution Quality (Optimal Cardinality) Analysis along with Response Time Analysis with varying number of Ants. The Number of Ants (NOA) taken for 25 and 30 joins is 1–25. For Number of Joins (NOJ) 35, 40, 45, 50, and 55, *NOA* is set to 1–30. Observations state that when the number of ants is extremely less, i.e., one to three, the algorithm tends to fall into local optimum exhibiting premature convergence. After achieving an optimal value, there is no significant improvement in Cardinality immaterial of the increased number of ants. The representation of Response Time is in seconds.

Table 2 Cardinality analysis and response time analysis with varying ants

Cardinalty analysis								Response time analysis						
Number of joins (NOJ)								Number of joins (NOJ)						
NOA	25	30	35	40	45	50	55	25	30	35	40	45	50	55
1	33	42	68	77	147	191	215	0.055	0.102	0.134	0.203	0.223	0.365	0.363
2	30	39	58	69	102	198	199	0.135	0.162	0.23	0.313	0.415	0.789	0.704
3	35	46	47	56	71	163	183	0.161	0.217	0.303	0.447	0.62	0.818	1.037
4	26	41	36	45	87	159	191	0.191	0.352	0.418	0.797	0.813	1.081	1.386
5	27	33	38	52	82	172	185	0.250	0.434	0.526	0.733	1.2	1.334	2.228
6	**24**	36	36	40	57	168	182	**0.270**	0.42	0.616	0.881	1.213	1.623	2.051
7	26	32	45	52	85	159	182	0.344	0.483	0.712	1.022	1.393	1.589	2.465
8	29	**30**	37	40	60	163	190	0.382	**0.555**	0.802	1.499	1.984	2.533	2.814
9	25	37	37	52	55	172	186	0.392	0.612	0.919	1.300	1.781	2.273	3.081
10	24	30	39	45	53	157	188	0.429	0.864	1.341	1.491	2.002	2.698	3.145
11	31	34	**34**	40	59	163	194	0.473	0.743	**1.083**	1.563	2.917	3.636	3.752
12	25	30	37	41	**45**	**153**	192	0.523	0.840	1.214	2.184	**2.374**	**3.142**	5.004
13	34	30	36	**39**	47	154	187	0.545	1.115	1.308	**1.867**	3.355	3.243	4.573
14	27	33	39	40	57	158	180	0.572	0.941	1.735	2.039	2.568	4.036	4.784
15	27	36	39	40	46	172	183	0.625	1.003	1.502	2.736	2.970	3.289	5.193
16	29	35	35	41	44	173	**177**	0.685	1.084	1.576	2.283	3.159	4.212	**5.454**
17	26	34	37	39	54	172	191	0.841	1.141	1.701	2.405	3.36	4.447	5.786
18	28	37	36	39	46	164	179	0.754	1.215	2.155	2.620	3.563	4.712	6.136
19	29	34	37	41	49	168	189	0.795	1.289	2.34	2.261	3.729	4.931	6.422
20	29	31	38	39	44	153	178	1.093	1.412	2.018	2.941	3.900	5.266	6.687
21	25	32	35	43	45	160	186	0.885	1.385	2.100	2.972	4.106	5.488	8.730
22	26	39	34	40	49	169	170	0.965	1.658	2.919	3.123	4.278	6.188	7.591
23	27	33	37	40	48	164	183	1.011	1.991	2.247	3.24	4.466	6.338	7.893
24	25	35	39	39	45	167	175	1.658	2.650	2.313	3.403	4.699	6.297	8.182
25	24	36	38	42	47	165	188	2.031	3.010	2.759	3.893	5.059	8.199	8.428
26			37	41	45	162	178			3.214	4.651	5.04	6.777	8.864
27			39	39	50	172	181			3.971	4.860	5.297	7.160	9.143
28			33	45	51	168	185			4.719	5.629	5.365	8.047	11.376
29			36	43	56	162	194			5.314	6.010	5.435	8.012	10.022
30			38	39	49	150	175			6.05	6.910	5.655	8.236	10.596

drksujatha23@gmail.com

Table 3 Minimum ants analysis optimal solutions

NOJ	Cardinality when NOA = 1	Min NOA	Optimized cardinality for min NOA	IP (%)	Ants ratio
25	33	6	24	27.27	0.24
30	42	8	30	28.57	0.26
35	68	11	34	50.00	0.31
40	77	13	39	49.35	0.32
45	147	12	44	70.07	0.26
50	191	12	153	17.80	0.24
55	215	16	177	17.67	0.29

When the number of ants is extremely low, minimum time is taken to generate optimized result sets. As the number of ants increases, the random wandering of the ants increases due to lack of pheromone on trial paths at an initial stage leading to increased Response Time [11, 18]. The analysis also indicates that as the number of ant increases, the Cardinality decreases. It is essential to estimate the minimum number of ants needed to generate effective optimal solutions with considerable response time. Table 3 displays the analysis of the Solution Quality defined in terms of Join Cardinality.

Improvement Percentage (IP) estimates the improvement in Cardinality of the join operation when number of ants increases from one to estimation of optimized results. The Improvement Percentage that is evaluated using the Eq. [1] is given below

$$IP = \frac{Max\,Tuples\,(when\,NOA = 1) - Optimal\,Tuples}{Max\,Tuples} * 100 \tag{1}$$

The results in Table 3 show that as the number of Joins changes in the distributed query, the number of ants needed to generate optimal results also changes. Hence, this experiment attempts to create artificial ants dynamically depending on the number of Joins in the underlying input query. To perform this task, ***Ant Ratio*** is coined as ***Ratio of Ants: Joins***. It is defined as the minimum number of ants needed to generate the optimal result with respect to the number of Joins involved in distributed query. Equation [2] defines the Ant Ratio as

$$AntRatio = \frac{Mimimum\,Number\,of\,Ants}{Number\,of\,Joins} \tag{2}$$

Observations on evaluation of Ant Ratio states that optimal results are obtained when the ratio falls in the range 0.2–0.3. In other words, to generate optimized results of the query, the optimal number of ants should be 1/5th–3/10th of the problem size. Figure 1 displays the graphical representation of the Ant Ratio Analysis against the number of Joins. The horizontal axis represents the Number of Joins while the vertical axis represents the Ant Ratio as evaluated in Table 3.

drksujatha23@gmail.com

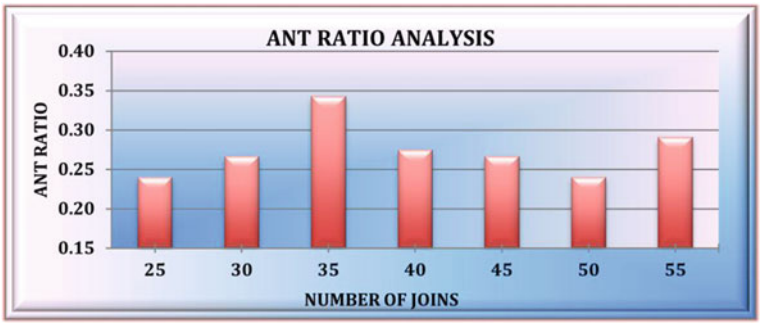


Fig. 1 Ant Ratio versus Number of Joins

The following observations signify the Number of Ants (NOA) needed to optimize fragmented query from above experiments:

1. ***1 >=NOA >=5***: When the Number of Ants is in the range of 1–5, Cardinality is extremely high and the execution time is less showing a premature convergence.
2. ***6 >=NOA >=15***: Optimal results are generated when the number of ants falls in this range with Ant Ratio ***0.2 >= Ant Ratio >= 0.3***.
3. ***NOA >=16***: With large number of ants, the Query Response Time increases due to high level of exploration at an initial stage. This leads to slow convergence speed.

5 Conclusion

This paper focuses on the estimation of minimum number of ants needed to optimize distributed query using Ant Colony Optimization Algorithm in an adaptive manner. The query result analysis is in terms of both Join Cardinality and Response Time. Experimental results show that as the number of ant increases, the execution time also increases because of the unsystematic exploration of the ants at an initial stage due to negligible amount of pheromones. However, as the number of ants increases, the Join Cardinality improves. To estimate minimum number of ants needed to generate optimized results, Ant Ratio is calculated. It emphasizes on the minimum number of ants needed to generate optimized solutions fall in the range of 0.2–0.3, i.e., the minimum number of ants needed is in the range of 1/5th–3/10th of the problem size.

drksujatha23@gmail.com

References

1. Adel A, Ahmadzadeh M (2013) The optimization of running queries in relational databases using ant-colony algorithm. Int J Database Manag Syst (IJDMS) 5(5)
2. Alamery M, Faraahi A, Javadi HH, Nourossan S (2010) Multi-join query optimization using the bees algorithm. Adv Intell Soft Comput 449–457
3. Ceri S, Pologatti G (1984) Distributed database principles and systems. McGrawHill Publication
4. Colorni A, Dorigo M, Maniezzo V (1994) Ant system for job-shop scheduling. Belg J Oper Res Stat Comput-Sci 34(1):39–54
5. Connolly T, Begg C (2007) Database systems-a practical approach to design, implementation and management 3rd ed. Pearson Education
6. Dorigo M, Gambardella L (1997) Ant colony system: a cooperative learning approach to the traveling salesman problem. IEEE Trans Evol Comput 1(1):53–66
7. Dorigo M, Stutzle T (2005) The ant colony optimization metaheuristic: algorithms, applications, and advances. Handbook of metaheuristics international series in operations research & management science
8. Dorigo M, Mauro B, St¨utzle T (2006) Ant colony optimization-artificial ants as a computational intelligence technique. IEEE Comput Intell Mag
9. Duan P, Yong A (2016) Research on an improved ant colony optimization algorithm and its application. Int J Hybrid Inf Technol 9(4):223–234
10. Gambardella LM, Taillard E, Dorigo M (1999) Ant colonies for the quadratic assignment problem. J Oper Res Soc 50(1):167–176
11. Golshanara L, Rouhani R, Shah-Hosseini H (2014) A multi-colony ant algorithm for optimizing join queries in distributed database systems. Knowl Inf Syst 39:175–206
12. Kadkhodaei H, Mahmoudi F (2011) A Combination Method for Join Ordering Problem in Relational Databases using Genetic Algorithm and Ant Colony. *IEEE International Conference on Granular Computing*
13. Kumar T, Singh R, Kumar A (2015) Distributed query plan generation using ant colony optimization. Int J Appl Metaheuristic Comput 6(1):1–22
14. Li N, Liu Y, Dong Y, Gu J (2008) Application of ant colony optimization algorithm to multi-join query optimization. Adv Comput Intell Lect Notes Comput Sci 189–197
15. Liu C, Yu C (1993) Performance issues in distributed query processing. IEEE Trans Parallel Distrib Syst 4(8)
16. Ozsu MTM, Valduriez P (1999) Principles of distributed database systems 2nd edn. Princeton-Hall, Englewood Cliffs, NJ
17. Shweta, Singh A (2013) An effect and analysis of parameter on ant colony optimization for solving travelling salesman problem. Int J Comput Sci Mob Comput 2(1):222–229
18. Wagh A, Nemade V (2017) Query optimization using modified ant colony algorithm. Int J Comput Appl 167(2):29–33
19. Wang C, Chen M (1996) On complexity of distributed query optimization. IEEE Trans Knowl Data Eng 8(4)
20. Wei X (2014) Parameter analysis of basic ant colony optimization algorithm in TSP. Int J u-and e-service Technol 7(4):159–170
21. Yu C, Chang C (1984) Distributed query processing. ACM Comput Surv 16(4):399–433

drksujatha23@gmail.com

Comparative Study of Various Cryptographic Algorithms Used for Text, Image, and Video



Nilesh Advani, Chetan Rathod and Atul M. Gonsai

Abstract In today's era of computer, the most important aspect for data is its security. This is very much essential when the data is traveling through various mechanisms of wired and wireless communication. We authors, tend to do comparative study on various techniques (algorithms) which are available for securing the data in terms of encryption and decryption. In this paper, we have studied various research papers of comparisons and based on that concluded that preliminarily which encryption algorithm would be better for our future research. We have tried to do literature review research and comparative analysis on AES, DES, 3DES, Blowfish, Threefish, RC variations, A5, DH, DSS, Elliptic Curve, RSA, DSA, etc.

Keywords Security · Encryption algorithms · Symmetric · Asymmetric
AES · DES · 3DES · Blowfish · Threefish · RC · A5 · DH · DSS
ECDH · RSA · DSA

1 Introduction

There are many **symmetric and asymmetric encryption techniques**, which are used for encryption and decryption purposes. Symmetric works on common key, whereas Asymmetric works with one public and one private key. We refer to following diagram Fig. 1.

N. Advani (✉) · A. M. Gonsai
Department of Computer Science, Saurashtra University, Rajkot, Gujarat, India
e-mail: nileshadvani@gmail.com

C. Rathod
Vivekanand College for Advance Computer & Information Science,
VNSGU, Surat, India
e-mail: rathodchetan@yahoo.co.in

© Springer Nature Singapore Pte Ltd. 2019

V. S. Rathore et al. (eds.), *Emerging Trends in Expert Applications and Security*,
Advances in Intelligent Systems and Computing 841,
https://doi.org/10.1007/978-981-13-2285-3_46

drksujatha23@gmail.com

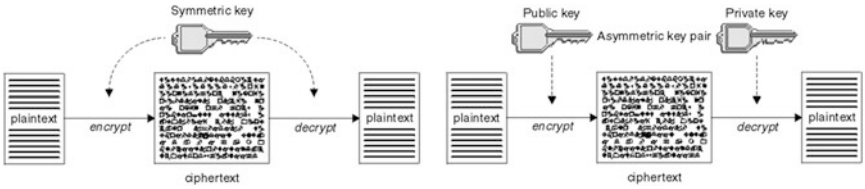


Fig. 1 Working of public & private keys

There are many types of Symmetric and Asymmetric Algorithms available as follows:

Symmetric key algorithms	AES (Advanced Encryption Standard) [1], DES (Data Encryption Algorithim) [1] variations 3DES, DECX., Blowfish, TwoFish, ThreeFish, RC (Rivest Cipher) variations, RC2, RC4, RC5, RC6, A5/1, A5/2 (mainly for GSM), Serpent, Skipjack, IDEA, Chacha
Asymmetric key algorithms	Diffie–Hellman, DSS (Digital Signature Standard), Elliptic Curve, RSA (Rivest–Shamir–Adleman), ElGamal (based on Diffie–Hellman), DSA (Digital Signature Algorithm)

Many researchers have contributed towards proving best suitability of various asymmetric and symmetric algorithms. They have shown comparisons of both types of algorithms by testing them in terms of encryption–decryption time, CPU execution time, Memory Utilization, Secrecy, possibility to crack the same, etc. From our research, we found that some researchers have contributed towards Fully Layered and Selective Encryption [2] methods also where they encrypt only selected portions of the data.

The objective of research is to come at the point to decide which algorithm shall be used (Asymmetric or Symmetric) and whether to use Block Cipher or Stream Cipher. After the above things are proved, the particular design of algorithm will be done using suitable tool and then it will be executed under live environment.

2 Various Encryption Algorithms

As discussed earlier, encryption algorithms are used for cryptography of Video, Audio, Text, Image, etc. All the algorithms use key to encrypt and decrypt the data in various ways. Some important encryption algorithms are given as follows:

A. DES (Data Encryption Standard) and 3DES (Triple Data Encryption Standard)
Initially, DES was developed which was using 56 bits of key size and was having only one round of encryption [3]. Aarti Devi, Ankush Sharma, and Anamika Rangra say that DES was not secure or you can say that most corrupted algorithm in

drksujatha23@gmail.com

the history of encryption algorithms. It was mainly designed hardware, whereas for software, it gives less performance.

On the counterpart, Triple DES uses three individual keys with 56 bits each. Hence, the total key length goes to 168 bits. Some authors [4] also say that 3DES uses 112 bits as a key.

B. AES (Advanced Encryption Standard)

As per details given by AL. Jeeva and Dr. V. Palanisaamy, the encryption ration for AES are high. It is proved efficient in 128-bit form. AES also uses keys of 192 and 256 bits for heavy-duty encryption purposes. The speed of AES is high.

AES works in different rounds of 10, 12, and 14 for encryption. Hence, as per many authors, this is treated as one of the best algorithms so far. As per authors, it is suitable for both hardware as well as software.

C. Blowfish

Blowfish is yet another algorithm, which was mainly designed to replace DES. This algorithm splits messages into blocks of 64 bits and encrypts all the messages individually. As per some authors, blowfish goes for 14 rounds [5] and some authors say that blowfish goes for 16 rounds [2, 6].

The key length of Blowfish is not fixed. It can range between 32 and 448 bits. One of the good advantages of using Blowfish is, it gives security against [7] Dictionary attacks. As per some authors [8], blowfish is good algorithm amongst AES, DES variations.

D. RSA (Rivest—Shamir—Adleman)

This is one of the initial algorithms, which works on public encryption key and secret decryption key. As per authors [4], RSA is the most popular encryption technique used till now. RSA uses 1024 bit streams for keying purpose which today can also be up to 4096 bytes typically [9]. The main thing about RSA is it depends on large prime integer numbers. However, in the past, some authors [7, 10] have also specified about various types of attacks, which can occur over RSA.

E. RC Variations (RC2, RC4, and RC6)

RC is a series of symmetric encryption algorithms developed by RSA security. Many authors write that there is a range of RC from RC1 to RC6. But mainly authors talk about only RC2, RC4, and RC6. RC2 utilizes key size between 8 and 1024 bits and utilizes block size of 64 bits. RC4 works for key size between 40 and 2048 bits. It was leaked on 1994. RC5 has a variable block size (32, 64, or 128 bits), key size (0–2040 bits), and number of rounds (0–255). The original suggested choice of parameters was a block size of 64 bits, a 128-bit key, and 12 rounds. RC6 is derived from RC5, which works on key sizes of 128, 192, and 256 bits, whereas works on 128 bits block size.

drksujatha23@gmail.com

Some authors [11] have developed RC6e (Enhancement of RC6) which works on block size of 256 bits, whereas Modified RC6 (MRC6) works on block size of 512 bits.

F. A5/1 and A5/2

This is stream cipher algorithm, which is used for GSM technology. A5/1 and A5/2 are one of the highly secure algorithms, which are used for [12] GSM standard used for Wireless Communication Medium. These are stream cipher algorithms, hence, there is no need to specify on how much block size they work for. Some authors [13] have invented enhanced version of A5/1 and modified the original algorithm to prove better for Image encryptions. Along with this, some authors have also developed A5 algorithm [14] with sponge techniques, which can be discussed on subsequent papers if required.

G. Twofish and Threefish

These two algorithms are developed specially by keeping Internet security in mind. [15] Twofish works on 128-bit block cipher with various key lengths of 128 bits, 192 bits, and 256 bits [16], whereas Threefish [17] works as tweakable block cipher with block sizes of 256 bits, 512 bits, and 1024 bits. This works on fixed key size of 128 bits. Threefish algorithm uses XOR and modulus addition for good security. It was proved as less secure and slow [18] in comparison to AES.

H. ECDH (Elliptic-Curve Diffie–Hellman)

This is asymmetric cryptography technique, which requires smaller keys in comparison to other non-ECC cryptography. ECDH uses public–private key pair to establish a shared secret key for insecure channel. Actually, ECDH is an improvement over DH. The main advantages of ECC variations are it provides good security in small key sizes [19]. ECC can work on HexaDecimal data, which also works for ASCII data.

3 Comparative Analysis of Encryption Algorithms

Based on various encryption algorithms, many authors [5, 7] and researchers have proposed various types of comparison for these algorithms. In this paper, we have tried to combine all those comparisons where we have compared various factors of various algorithms in form of Key Length, Cipher Type, Block Size, Resistance, Possible Keys, Rounds, etc.

Table 1 is prepared in a separate Word file in Landscape format. Most of the references here are taken from [5, 7, 9, 16, 17].

drksujatha23@gmail.com

Table 1 Comparison table of various encryption algorithms

Factors	Algorithm							
	DES/3DES	AES	Blowfish	RSA	RC variations	A5 variations	Two–threefish	ECDH
Cipher type	Block (Symmetric)	Block (Symmetric)	Block (Symmetric)	Block/Stream (Asymmetric)	Block (RC4 Stream) (Symmetric)	Stream	Block (Symmetric)	Block (Asymmetric)
Key length	56 bits (DES), 112/168	128, 192, 256	32–448	1024 to 4096	(8–128 bits in RC2) (128, 192 and 256 bits in RC6)	Not applicable	(128, 192, 256 Twofish) (256, 512, 1024 Threefish)	Key exchange Management
Block size	64 bits (both)	128, 192, 256	64 bits	Depends on key length 86–1024, 214–2048 [9]	64 bit (RC4), 128 bits (RC6)	Not applicable	128 (Twofish), (256, 512, 1024 Threefish)	–
Encryption ratio	High (DES), Moderate (3DES)	High	High	HIGH	Low (RC4), High (Other variations)	High	High	High
General security	3DES Ok but DS proved inadequate	Considered secure	Vulnerable	Considered secure	Vulnerable	Considered secure	Less in comparison to AES	Not Specified
Security against attacks	Brute force, chosen and known text (3DES)	Chosen and known plain text	Dictionary attacks	Timing attacks	Bit flipping	Voice privacy	Breaks 6 rounds out of 16 of 256 bit key (Twofish)	Eavesdropping

drksujatha23@gmail.com

4 Conclusion

We have tried to compare most of the existing encryption algorithms with various types of parameters given by various researchers and authors. From the table given in above section, we can conclude that encryption ration is high in most of Symmetric algorithms. If we are willing to develop Stream Cipher, then it is better to use A5 variations. RSA seems to be better in Asymmetric algorithm, whereas AES seems better in Symmetric algorithm. Blowfish, RC variations were found vulnerable in comparison to others where as Twofish and Threefish were found less secure in comparison to AES.

References

1. Bhat B, Ali AW, Gupta A (2015) DES and AES performance evaluation. IEEE—Int Conf Comput Commun Autom ICCCA 2015:887–890
2. Gupta S, Kishor L, Goyal D (2014) Comparative analysis of encrypted video streaming in cloud network, Springer—nt J Comput Sci Inf Technol 5(4):5470–5476
3. Devi A, Sharma A, Rangra A (2015) A review on DES, AES and blowfish for image encryption & decryption. IJCSIT Int J Comput Sci Inf Technol 6(3):3034–3036
4. Iyer SC, Sedamkar RR, Gupta S (2016) A novel idea on multimedia encryption using hybrid crypto approach. Sci/Elsevier—Procedia Comput Sci 79:293–298
5. Mathur M (2013) Comparison between DES, 3DES, RC2, RC6, BLOWFISH and AES. In: Proceedings of the national conference on new horizons IT, pp 143–148
6. Bala M, Kumari P, Sharma A (2017) Comparative analysis of symmetric key algorithms: DES, AES and blowfish for audio encryption and decryption. IEEE, pp 1048–1054
7. Jeeva AL, Palanisamy DV, Kanagaram K (2012) Comparative analysis of performance efficiency and security measures of some encryption algorithms. Int J Eng Res Appl 2 (3):2248–9622
8. Tamimi AA (2008) Performance analysis of data encryption algorithms. IEEE retrieved Oct vol 1, pp 399–403
9. Keren S (2016) Typical block size in RSA. Cryptography—Stack Exchange. https://crypto.stackexchange.com/questions/32692/what-is-the-typical-block-size-in-rsa
10. Zulkarnain S, Idrus S, Aljunid SA, Asi SM, Sudin S, Ahmad RB (2008) Performance analysis of encryption algorithms text length size on web browsers. IJCSNS 8(1):20–25
11. Aggarwal K (2015) Comparison of RC6, modified RC6 & enhancement of RC6. In: IEEE Proceedings of 2015 international conference on advances in computer engineering and applications. ICACEA 2015, pp 444–449
12. Cattaneo G, de Maio G, Faruolo P, Ferraro Petrillo UF (2013) A review of security attacks on the GSM standard. In: Springer—IFIP international federation for information process, pp 507–512
13. Naveen C (2016) Image encryption technique using improved A5/1 cipher on image bitplanes for wireless data security. In: IEEE 2016 international conference on microelectronics computer communications
14. Wahab HBA, Mohammed MA (2015) Improvement A5/1 encryption algorithm based on sponge techniques. IEEE-2015 World Congr Inf Technol Comput Appl WCITCA 2015:2–6
15. Gandhi RA (2015) A study on current scenario of audio encryption. Int J Comput Appl 116 (7):13–17
16. Wikipedia (2017) Twofish—Wikipedia, Wikipedia. https://en.wikipedia.org/wiki/Twofish

drksujatha23@gmail.com

17. Wikipedia, Threefish—Wikipedia, Wikipedia. https://en.wikipedia.org/wiki/Threefish#Security
18. Rizvi SAM, Hussain SZ, Wadhwa N (2011) Performance analysis of AES and Twofish encryption schemes. In: IEEE, Proceedings-2011 international conference on communication systems network technologies CSNT 2011, pp 76–79
19. Keerthi DBSK (2017) Elliptic curve cryptography for secured text encryption. In: IEEE 3rd international conference on collaboration and internet computing

drksujatha23@gmail.com

A Comparative Study of Ontology Building Tools for Contextual Information Retrieval



Ripal Ranpara, Asifkhan Yusufzai and C. K. Kumbharana

Abstract After the evolution of semantic web technologies, ontology's play the key role in Knowledge representation, Knowledge Management, and information retrieval. Ontology can be defined as the grammar that can be interpreted by the machine. In short, it is defined as concept or term, which is used to represent the Knowledge. The aim of this paper is to give the comparative study of different ontology building and management tool (Protégé 4.3, Swoop, Apollo, and IsaViz) which are open source. In this paper, we will compare the mentioned tools in context of Cross-Platform Integration, Easy to update and manage, Tolerance, etc. As we are talking about World Wide Web, the diversity plays the major role as this tool can be used by the different groups of people. The mentioned tools cannot interchange the ontology's. This research also identifies the common feature between each tool and each tool has its own significance importance, so based on that this paper will represent the comparative study of different tools.

Keywords Ontology building tools · Protégé 4.3 · SWOOP · IsaViz
Apollo · Ontology · Semantic web

R. Ranpara · C. K. Kumbharana
Shree M.N Virani Science College, Atmiya Group of Institutions, Rajkot, India
e-mail: ranpararipal@gmail.com

C. K. Kumbharana
e-mail: ckkumbharana@yahoo.com

A. Yusufzai (✉) · C. K. Kumbharana
Department of Computer Science, TNRAO College,
Saurashtra University, Rajkot, India
e-mail: asyusufzai@yahoo.com

© Springer Nature Singapore Pte Ltd. 2019

V. S. Rathore et al. (eds.), *Emerging Trends in Expert Applications and Security*,
Advances in Intelligent Systems and Computing 841,
https://doi.org/10.1007/978-981-13-2285-3_47

drksujatha23@gmail.com

1 Introduction

The current era of the web is Semantic Web [1]. The Semantic Web [1] is the Web that can be interpreted by the machine and can retrieve the query based on the context of the user. The grammar that is interpreted by the hyperlinks and can understand the user context intelligently using a mechanism called ontology [2]. This concept can be defined as the set of all concept with the grammar and the relationship among different entity. Nowadays, ontology is used to represent the interrelatedness among different entities and the relationship among them. The main use of the ontology is information interpretation with context and reorientation of that information. The ontology creation and management requires many advance tools [3]. Ontology's are platform independent and cross-platform operable as it is a plain text grammar representation. Currently, there are many ontology development and management tools available like Protégé 4.3, IsaViz, Swoop, and Apollo [4–8].

2 Ontology Quality Evaluation Architecture

The most common and appropriate framework for the ontology quality evaluation is the Semiotic Framework [9, 10]. This paper referred to the framework and we have tried to describe it in our own research context. The framework is given by and can be evaluated by [11]. This paper also described the very short outline of the mentioned framework. The framework is focused on the Knowledge representation and inter-relatedness among the different node, which includes different objects like the human and tools. In short, the framework gives the quality of interdependent relationship between different objects, e.g., the word Apple can be interpreted as fruit also or it can be interpreted as Apple Company also. So accordingly, it will identify the relationship quality. The quality of the model can be examined using belowmentioned parameters.

3 Ontology Building Tools

In this section, we have represented the study of different open source ontology building tool. Ontology building or editor tools allows the user with Graphical user interface to create, manage, inspect, and browse the code of the ontology [12].

3.1 Protégé 4.3

Protégé 4.3 [5] is an editor and management tool for the ontology with Graphical User Interface. Using this tool, the user can create the ontology's of any domain

drksujatha23@gmail.com

with metadata grammar. Protégé 4.3 [5] In short an ontology and knowledge base editor produced by Stanford University. Protégé is a tool that enables the construction of domain ontology's, customized data entry forms to enter data. Protégé allows the definition of classes, class hierarchies, variables, variable value restrictions, and the relationships between classes and the properties of these relationships. Protégé is free and can be downloaded from http://protégé.stanford.edu [5].

3.2 Isaviz

IsaViz [6] is a visual environment for browsing and authoring RDF models as graphs. This tool is offered by W3C Consortium. IsaViz [6] was developed by Emmanuel Pietriga. The first version was developed in collaboration with Xerox Research Centre Europe, which also contributed with XVTM, the ancestor of ZVTM (Zoomable Visual Transformation Machine) upon which IsaViz is built. As of October 2004, further developments are handled by INRIA Futures project In Situ.

3.3 Apollo

Apollo [7] is a user-friendly knowledge management and representation editor. Internal model is built as a frame system according to the internal model of the OKBC protocol. Apollo currently does not support non-template class slots. For each class, it is possible to create many instances. An instance inherits all slots of the class. Each slot has a set of facets [7].

3.4 Swoop

Swoop [8] is a web-based OWL ontology editor and browser [3]. SWOOPs interface has hyperlinked competences so that it is very easy for the user to edit and build it again. SWOOP does not follow a methodology for ontology construction. Users can reuse external ontological data [3].

4 Determination of Various Parameters for Model Comparison

There are various parameters to measure the quality of the ontology building tools. According to the semiotic standard framework, the quality can be measured by the belowmentioned parameters.

drksujatha23@gmail.com

5 Creation of Ontology

To determine the success of the abovementioned ontology tools, we have created the two different cases, i.e., (1) Apple (Fruit) and (2) University domain. But to create the ontology for these two domain, we have followed the following steps.

5.1 Creation of Ontology

Step 1. The first to create ontology is to determine and finalize the domain of the ontology for which domain.
Step 2. The reusability of the grammar is the most useful feature of any ontology so reusing existing ontologies can be considered as a very useful step for the implementation.
Step 3. The third step is you need to identify and analyze the important terms in your ontology so that we can perfectly define the grammar of the same.
Step 4. The fourth step is the most important step in the ontology development so for that, we need to define the classes and we need to define the class hierarchy of the ontology.
Step 5. After defining the class and hierarchy of the class, we need to define the properties of classes for the better representation of the grammar ontology.
Step 6. The more elaborative step is the step six that define the Define the facets of the main classes.
Step 7. Finally, after defining the classes' properties, we need to create the instances of the same.

We have apply the above steps on the two domain for all the ontology tools And then, we have analyzed the following results shown in the below Table 1 in which we have considered all the dimensions given in the framework, and we have identified some facts regarding the different tools which is described in the Sect. 6.

Table 1 Ontology tools terminology

Functionality	Description
Create	It allows you to create your own resource description format using graphical user interface
Manage	It allows you to manage your ontology resource description format. It also allows you to modify the already developed ontology
Inspect	Using advance plugin and built-in features it allows you to debug your grammar using graphical user interface
Code	It automatically creates the code (rdf) of your ontology and you can browse the code using code editor of the tool

drksujatha23@gmail.com

6 Comparative Study and Analysis of the Abovementioned Tools

After creation of the abovementioned ontology, we have started the comparative study by considering the mentioned framework features (e.g., Physical, Empirical, syntactic, contextual, pragmatic, and social) which can be used to analyze the quality of the mentioned tools. And each feature is divided into the sub feature; each feature has its own significant importance. Each sub feature is implemented using tools and we have specified that experimental result in the given below Table 2. Each feature for the particular tools represents the quality of the ontology development tool. We have appraised the tools based on the major quality physical, empirical, Knowledge, Syntactic, Contextual, and Pragmatic features as described below:

- **Physical**: Physical feature is all about the extensibility of the tools. It can be measured based on the external and internal factor of the participant.
- **Empirical**: The empirical quality is the qualitative measure of any service, which is very difficult to measure. It is mainly deals with interoperability of the resource description format.
- **Syntactic**: Syntactic feature deals with syntactical debugging of the resource description format. It also measures the user-friendliness of the module.
- **Contextual**: It mainly deals with semantic contextual quality. It allows the user with tutorial and tooltip based on the grammar of the resource description format.
- **Pragmatic**: This quality represents the graphic visualization feature of the tool.
- **Social**: This feature is the independent feature, which is based upon the ethnography of the user.

The Table 3 shows the different supported parameters and its implementation outcome in the form of yes or no. The same is also represented using the graph as shown in Fig. 1. In which, small bar represents no and big bar represents yes.

Table 2 Framework features

Model quality	Language quality
Physical	Participant knowledge
Empirical	Social interpretation
Syntactic	Technical interpretation
Contextual	Domain modeling
Perceived	Comprehensive appropriateness
Pragmatic	
Social	
Knowledge	

drksujatha23@gmail.com

Table 3 Overall features analysis of different tools

Quality	Features	Apollo	IsaViz	Protégé 4.3	Swoop
Physical features	Extensibility	No	No	Yes	No
	Ontology storage space	Yes	Yes	Yes	Yes
	Graphical grammar taxonomy	No	Yes	Yes	Yes
	Versioning	No	No	Yes	No
	Export–import format pattern	No	No	Yes	Yes
Empirical features	Merging patterns	No	No	Yes	Yes
	Zooming features of visual	No	Yes	Yes	No
Knowledge features	Collaboration	No	No	Yes	No
	Advance library support	Yes	Yes	Yes	No
	Knowledge model	No	Yes	Yes	No
Syntactic features	Error checking	No	Yes	Yes	Yes
	Consistency checking	Yes	No	Yes	Yes
	Model reuse	Yes	Yes	Yes	Yes
Contextual features	Tutorial	No	Yes	Yes	Yes
	Tooltips	Yes	No	Yes	No
Pragmatic features	Visualization	Yes	Yes	Yes	Yes
	Filtration of objects	No	Yes	Yes	Yes

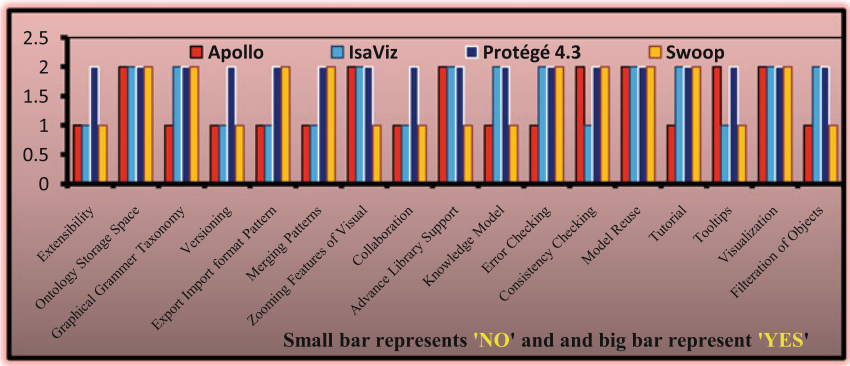


Fig. 1 Comparative study chart based on the framework

To determine the more suitability for the taken two ontologies, we have created the aggregated analysis that tells that how many feature is supported and how many feature is not supported by the different tools that is represented in the form of yes

drksujatha23@gmail.com

Table 4 Cumulative analysis of tools

Tools	No	Yes	Total
Apollo	10	7	17
IsaVizviz	7	10	17
Protégé 4.3	0	17	17
Swoop	8	9	17

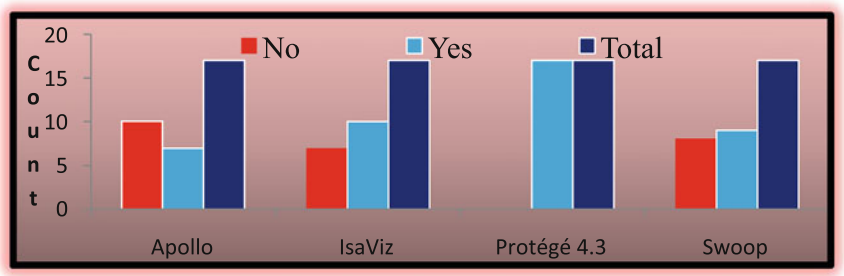


Fig. 2 Cumulative study chart based on the framework

and no based on the semiotic framework [9] as in Table 4, the same is also represented using the bar chart as shown in Fig. 2.

7 Conclusion

Ontology plays a vital role in the semantic web technology. So, the development and management of the ontology is the key concern. Many advance tools are the basic requirement for the abovementioned concern. Fortunately, there are many open source tools for the ontology development and management is available which is cross-platform operable. Ontology editor plays a major role in the management of the ontology. So, the overall research paper aims at comparing the open source ontology building tools available for the user, we have compared the tools based on operability, user-friendliness, import–export facility and patterns available for the development. In our paper, we have described the two different ontologies, which are created using different tools and we have analyzed that in our case the Protégé [5] tools is well appropriate with the output of 100% supported. IsaViz [6] tool is supporting 58.82%. The Swoop [8] is supporting 52.94%. Apollo [7] tool is supporting 41.18%. In our case, the inclusive study states that the protégé is highly compatible for developing the ontology with grammar, which is similar to our domain, i.e., (Apple (fruit) and university domain).

drksujatha23@gmail.com

References

1. Berners-Lee, T, Hendler J, Lassila O Article Title: The Semantic Web
2. Noy NF, McGuinness DL (2001) Ontology development 101: a guide to creating your first ontology. Stanford Knowl Syst Lab 25 (2001)
3. Kondylakis H, Flouris G, Plexousakis D (2009) Ontology and schema evolution in data integration: review and assessment. In: Proceedings of the confederated international conferences, CoopIS, DOA, IS, and ODBASE 2009 on the move to meaningful internet systems: part II, pp 932–947
4. Informatics SM (2010) The Protégé Ontology Editor and Knowledge Acquisition System welcome to protég. Knowl Acquis 2010–2010
5. Protege, "protege," Stanford Center for Biomedical Informatics Research at the Stanford University School of Medicine. https://protege.stanford.edu/
6. Pietriga E "IsaViz: a visual authoring tool for RDF," Aduna. https://www.w3.org/2001/11/IsaViz/
7. (UK) Knowledge Media Institute, Open University, "Apollo," Knowledge Media Institute, Open University, (UK). http://apollo.open.ac.uk/
8. Kalyanpur A, Parsia B, Sirin E, Grau BC, Hendler J (2006) Swoop: a web ontology editing browser. Web Semant 4(2):144–153
9. Su X, Ilebrekke L (2002) A comparative study of ontology languages and tools. In: Advanced Information Systems Engineering 14th International Conference on CAiSE 2002, pp 761–765
10. Krogstie GSJ, Lindland OI (1995) Defining quality aspects for conceptual models. In: Proceeding of the IFIP8.1 working conference on information systems concepts: towards a consolidation of views. Marburg, Germany
11. Goguen J (1999) Lecture notes in computer science, vol 1562, no 1562–1562 Chapter 15
12. Uschold M, King M (1995) Towards a methodology for building ontologies. Methodology 80 (July):275–280

drksujatha23@gmail.com

A Comparative Study of Cryptographic Algorithms for Cloud Security



Asifkhan Yusufzai, Ripal Ranpara, Mital Vora and C. K. Kumbharana

Abstract In the growing technology in the world, there has been a drastic change in the usage, popularity, and requirement of cloud computing. With the popularity of cloud computing, it's required and need arise to secure the data in cloud. In the cloud computing, security is a prime concern as well as it is also found as a critical issue. There is an increasing demand to secure data in cloud computing and usage of security algorithms. So, we have discussed in detail about cryptographic algorithms with different security issues and our main goal of this paper is to compare three cryptographic algorithms.

Keywords Cloud computing · Cryptographic algorithm · Cloud security

1 Introduction

In the last decade in the Information and Technology world, cloud computing has created demand and popularity. It has provided optimum level of resource utilization, new strategies with minimal costs. It has several drawbacks on different levels of implementation and vulnerabilities as well as security issues arise on

A. Yusufzai (✉) · R. Ranpara · M. Vora · C. K. Kumbharana
TNRAO College, Shree M.N Virani Science College, Rajkot, India
e-mail: asyusufzai@yahoo.com

R. Ranpara
e-mail: ranpararipal@gmail.com

M. Vora
e-mail: vora.mital@gmail.com

C. K. Kumbharana
e-mail: ckkumbharana@yahoo.com

A. Yusufzai · R. Ranpara · M. Vora · C. K. Kumbharana
Department of Computer Science, Saurashtra University, Rajkot, India

© Springer Nature Singapore Pte Ltd. 2019

V. S. Rathore et al. (eds.), *Emerging Trends in Expert Applications and Security*,
Advances in Intelligent Systems and Computing 841,
https://doi.org/10.1007/978-981-13-2285-3_48

drksujatha23@gmail.com

virtualization and multi-tenant. There are different cloud vendors and clients, which have implemented various services and different models for cloud computing [1].

We have investigated in the cloud security issues and management problem and found that there is a major problem arising in the cloud security management processes between cloud consumers and the cloud service providers. In detail, we have discussed the cloud computing basic three deployment models.

1. Private Cloud Model: It is implemented and managed by single person or single organization, whether managed solely or by a third party and hosted internally or externally. ***2. Public Cloud Model***: The public cloud is a most common deploying model, where resources like storage, service, and computing are offered by provider over the internet. ***3. Hybrid Cloud Model***: where two or more cloud (private, community, or public) allowing data and application to be shared between them. Basically, there are three types of services provided by cloud vendors. ***1. Infrastructure as a Service (IaaS)*** [2]: It provides the consumer to virtualize computing resources over the Internet. ***2. Platform as a Service (PaaS)*** [2]: It provides platform to customer for developing, executing and managing application. ***3. Software as a Service (SaaS)*** [2]: It allows user to use applications and cloud-based apps over the internet such as email, calendar, and office tools.

2 Security Issues and Challenges

A. ***Data Security Issues***: Data protection plays a vital role in cloud computing, however, it turns into a first-rate venture whilst Saas customers should depend upon their vendors for proper safety. The Saas provider is the only responsible for the information at the same time as is being processed or saved [3, 4].
B. ***Privacy Issues***: When data stored on cloud, there should be some security measure is that it is accessible only by that user with strict authentication process. It should be restricted for any unauthorized user to reach the sensitive data [3].
C. ***Data confidentiality Issues***: The cloud customers needs to make certain that their facts are kept exclusive to outsiders, along with the cloud company and their capability competitors [5].
D. ***Accessibility Issues***: Having access to programs over the internet via net browser makes access from any network device less difficult, which includes public computer systems and mobile devices. However, it additionally exposes the provider to extra safety dangers [5].

drksujatha23@gmail.com

3 Security Algorithm Used in Cloud Computing

Cloud computing uses Internet platform for communication over the network and encryption algorithms plays an important role to provide secure communication. The following three cryptographic algorithms are widely used on cloud computing for securing data.

3.1 RSA (Rivest–Shamir–Adleman)

RSA is asymmetric cryptographic algorithm. Karthija et al. [6] which is based on property of positive integers. RSA used modular exponential for encryption and decryption. Asymmetric actually approach that it really works on two different keys, i.e., public key and private key, which provide both authentication and security. Public key is used for encryption process and it provides security such that only the person with the exact private key can decrypt the chipper text. The public key may be regarded to everybody and it is also used for encrypting message. Message encrypted with public key can simplest be decrypted the usage of the private key. The process is shown in Fig. 1 [7].

RSA and DES agorithms which provides single level of encryption as well as decryption for data storage on cloud, which can very easily attacked and cracked.

3.2 Data Encryption Standard (DES)

In this algorithm, each block is uses a cipher secret key [8], which operate on plain block of 64 bits and also returns ciphertext block of same size. Data Encryption Standard (DES) which operates on 64-bit blocks where key sizes of 56 bits because

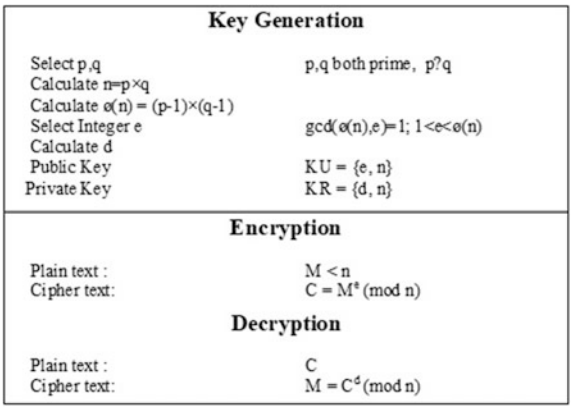


Fig. 1 RSA algorithm

drksujatha23@gmail.com

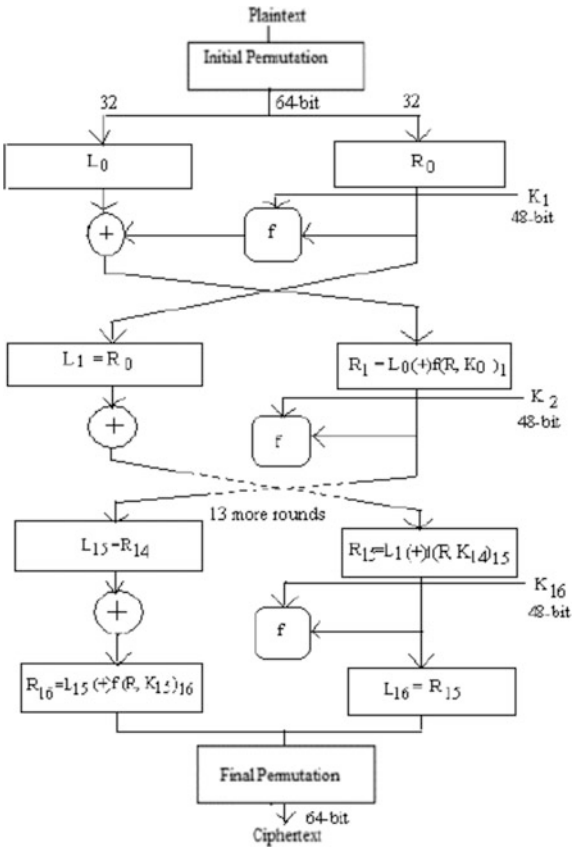


Fig. 2 Encryption with DES algorithm

every eighth bit in key like 8, 16, 24 … 56 are not used that is why actual key size is only 56 bits. In the encryption process, each 64 bits block is divided into 32–32 blocks. After completing each round, it uses a different 48-bit round key, which is generated by cipher key according to a algorithm as shown in Fig. 2 [9].

3.3 Advance Encryption Standard (AES)

This cryptographic algorithm has given by Belgian cryptographers "Joan Daemen" and "Vincent Rijmen" [10]. It is beneficial whilst we need to encrypt an exclusive text into a decrypt layout, best example of that when we are using sensitive or confidential data in email. The decryption of the encrypted text is only possible when we have the right password. AES is an iterative in place of Feistel cipher.

drksujatha23@gmail.com

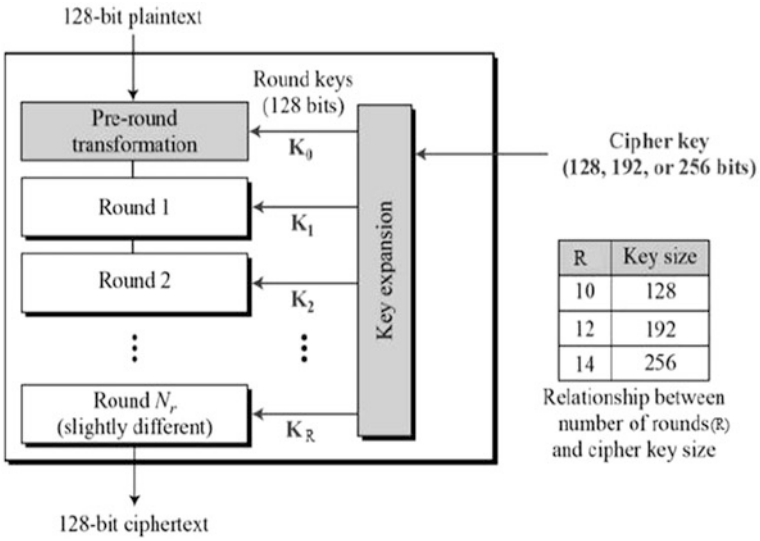


Fig. 3 The schematic of AES structure

AES is a symmetric block cipher that has been analyzed notably and is used broadly used for securing records on cloud.

AES is a block cipher, which contain block size of 128 bits. It allows 128, 192, or 256 bits key lengths [11]. AES performs continually well in each hardware and software program platforms beneath a wide variety of environments. This algorithm supports 8-bit and 64-bit platforms and DSPs. Statistical evaluation of the ciphertext has not been possible even after using huge range of check cases. It is used broadly with cloud information protection because of that there isn't any differential and linear cryptanalysis assaults were yet proved on AES. The schematic of AES structure is given in the following Fig. 3.

AES interesting features is that it performs all operations on byte instead of bits. Additionally, AES has variable key length and process of decryption is similar to the encryption process, each consist of four processes (Add round key, Mix column, Shift row, and Byte substitution) in reverse order.

4 Comparison of Existing Algorithms

We have seen that all three AES, DES, and RAS algorithms are having different features and parameters that are shown in Table 1 [12].

drksujatha23@gmail.com

Table 1 Cryptographic algorithm comparison with different parameters

Characteristic	RSA	DES	AES
Developed	1973	1977	2000
Block size	Minimum 512 bits	64 bits	128, 192, or 256 bits
Key length	>1024 bits	56 bits	128, 192, or 256 bits
Security	Least secure	Not secure enough	Excellent secured
Ciphering and deciphering key	Different	Same	Same
Algorithm	Asymmetric	Symmetric	Symmetric
Speed	Slower	Moderate	Faster

5 Conclusion

In this research paper, we have compared AES, DES, and RSA cryptographic algorithms and find out the best algorithms out of three, which can be used for Cloud data security. Cryptographic algorithms are used for securing data on Cloud, and we have made comparison of all three algorithms with its key component as well as different parameters and finally, we found that AES algorithm which takes less time to execute data on cloud. DES algorithm takes less time for encrypting data on cloud. RSA used large size of memory and longest time for encryption. In modern era, cloud computing is rapidly evolving model on Internet, there is a great deal about security issues at all levels, and although using encryption technique to secure data is not sufficient because it is difficult to find out that where actually data stored on cloud and how it can be accessed.

References

1. Gariba ZP, van der Poll JA (2017) Security failure trends of cloud computing. In: 2017 IEEE 3rd International Conference on Collaboration and Internet Computing, no Vm, pp 247–256
2. Majumder A, Roy S, Biswas S (2012) Data security issues and solutions in cloud computing. Int J Adv Res Comput Sci Softw Eng 2(10):212–231
3. Prakash GL, Prateek M, Singh I (2014) Data security algorithms for cloud storage system using cryptographic method, vol 5, no 3, pp 54–61
4. Bertino E (2016) Editorial: introduction to data security and privacy. Data Sci Eng 1(3): 125–126
5. S Kaur (2015) Survey of security algorithms in cloud, vol 115, no 19, pp 23–27
6. Karthija T, Radhamani AS, Vincy VGAG, Swaminathan LA (2017) An overview of security algorithms in cloud computing. Int J Recent Trends Eng Res 3(9):67–75
7. Reddy MR, Priyadarshini S, Karthikeyan B, Nadu T (2017) A modified cryptographic approach for securing distributed data storage in cloud, pp 131–139

drksujatha23@gmail.com

8. Khan SS, Tuteja PRR (2015) Security in cloud computing using cryptographic algorithms. Int J Innov Res Comput Commun Eng 3(1):148–154
9. Satyanarayana K (2016) Multilevel security for cloud computing using cryptography, vol 5, no 2, pp 235–242
10. Pancholi VR, Patel BP (2016) Enhancement of cloud computing security with secure data storage using AES. Int J Innov Res Sci Technol 2(9):18–21
11. Arora R, Parashar A (2013) Secure user data in cloud computing using encryption algorithms. Int J Eng Res Appl 3(4):1922–1926
12. Kaur R, Kinger S (2014) Analysis of security algorithms in cloud computing. Int J Appl Innov Eng Manag 3(3):171–176

drksujatha23@gmail.com

Mathematical Modelling and Analysis of Graphene Using Simulink Technique



Pragati Tripathi and Shabana Urooj

Abstract Silicon solar cell is now accepted as the better energy usage system according to the capital investment but we all know that it is not independently very much efficient. So far, it is simulated with graphene to enhance the output. The paper will show and study the factors like current density, absorption coefficient and wavelength spectrum of sun are studied by relating their characteristics and equation to each other to clear some dependency scenario of solar energy on graphene. Here, the equations are implemented by modulating the Simulink model and therefore, it is synthesized by Finite Difference Method using MATLAB tool, and the resulting graphs are obtained by SIMULINK toolbox through calibrating the parameters. The radiative recombination in GaAs tends to give loss of nearly $\sim 5\%$ of light generated carrier loss. Because of large optical absorption coefficient of GaAs, calculation of generation rate of carrier charges is done by not including light's multi-reflection in solar cell.

Keywords Renewable energy · Solar energy · Generation rate of electrons and holes · Graphene · Sun spectrum · Current density

Masterminded EasyChair and created the first stable version of this document.
Created the first draft of this document.

P. Tripathi (✉) · S. Urooj
Department of Electrical Engineering, Gautam Buddha University, Greater Noida, India
e-mail: pragati.knp022@gmail.com

S. Urooj
e-mail: shabanaurooj@ieee.org

© Springer Nature Singapore Pte Ltd. 2019

V. S. Rathore et al. (eds.), *Emerging Trends in Expert Applications and Security*,
Advances in Intelligent Systems and Computing 841,
https://doi.org/10.1007/978-981-13-2285-3_49

drksujatha23@gmail.com

1 Introduction

Days coming are relating the interest of the energy producers from non-renewable energy resources. Due to the duped methodology of renewable energy resources, it is a becoming a big substitute of conventionally used non-renewable energy resources. While surmising renewable energy as a whet for simulating the appetite of energy, many options appear like wave energy, solar energy, tidal energy, geothermal energy and many more all of them are pre-eminent from each other in there very different terminologies. The paper here will spurt towards the solar energy with infuser of graphene with it. Graphene is combined in this work, as solar energy harvesting technique is not so efficient if working in isolation. Very recent advances in graphene-based solar cells have seen the reflectance of solar rays reduced by 20%, which provides a potential efficiency increases up to 20. The grade of Graphene doping on the material, which is used as the base and the calculated quantity of layers of Graphene which are applied on solar module, these are two very strong potential parameters which affects the nature of the device [1].

The main limitation with Graphene is that it cannot be primarily accessible for research is that commercially the good quality of Graphene is very expensive or rarely available, the process is complex, and a monolayer of Graphene included toxic chemicals via exposing Platinum, Nickel or Titanium Carbide to ethylene or benzene at high temperatures [2]. This sternly restricted its application in electronics, as it was complicated, at that instance, to segregate graphene layers from its metallic substrate without damaging the Graphene. Graphene, a newly recognized material prepared from perforate layers of carbon is just one atom wide [1]. It is the lightest, robust and slimmest pre-eminent heat and energy conducting material yet obtained. We can use graphene where the light weighing application is used and it finds application in nanocomposite materials and it can conduct heat and electricity because both process include transportation of electrons [3]. The graphene installed on the solar panel is affected by the factors like solar power intensity, wavelength of suns spectrum, generation rate of electron and holes in the graphene layer [4]. As it is stated that the current is dependent on the electron flow in the wire which is affected by the current density of electrons and the polarity of the flow of the current in the system to flow in the circuit and it can also be termed as the polarizability of system of the upgraded solar panel [5]. The paper here is discussing the different characteristic of the factor at which the graphene efficiency is increasing. The factors, which are discussed in paper, are power intensity, No. of days, wavelength of sun spectrum, generation rate of electrons and holes, absorption coefficient angle, wavelength and current density.

drksujatha23@gmail.com

2 Results and Discussion

2.1 Power Intensity

The paper will do the systematic approach to obtain incident power within a certain wavelength interval ($\lambda_m - \Delta\lambda$, $\lambda_m + \Delta\lambda$) [6];

$$P_0(\lambda_m) = \int I_0(\lambda_m) d\lambda \tag{1}$$

where $I_0(\lambda_m)$ is the power intensity of sunlight with wavelength λ_m. Here, in the Eq. 1 for the intensity of light transferred per unit area is calculated by Electromagnetic radiation technique. The solar constant can be used to calculate the irradiance incident on a surface perpendicular to the sun's rays outside and the Earth's atmosphere on any day of the year from first to last day of the year. The Simulink Fig. 1a is showing the response of the solar power intensity with respect to the number of days in which N is denoting the day of the Year such that for January the 1st n =. The power intensity of Sunlight Wavelength is given by the given Eq. 2, here extraterrestrial irradiance on a plane perpendicular to the sun's rays is denoted by Io [7]

$$I_o = Isc\,[1 + 0.034\cos(360N/265.25)] \tag{2}$$

Here,
Isc = Solar constant = 1367 W/m^2.
Graph in Fig. 1b shows the variation in Io over the course of a year.

2.2 Rayleigh Scattering

When the light scatters off the molecular particle of air can be stated as the Rayleigh scattering. And any particle which is nearly one-tenth of the wavelength of the light can become a reason of scattering of light. The sky appears blue because of this phenomena. This is referred to Fig. 1c. Lord Rayleigh discovered this scattering phenomena and he used dipole scatterers which are very small in size as compared to the wavelength of light [8]:

$$I = I08\pi N\alpha\, 2/\lambda^4 R^2(1 + \mathrm{Cos}2\theta) \tag{3}$$

where

N # of scatterers
α Polarizability
R Distance from scatterer.

drksujatha23@gmail.com

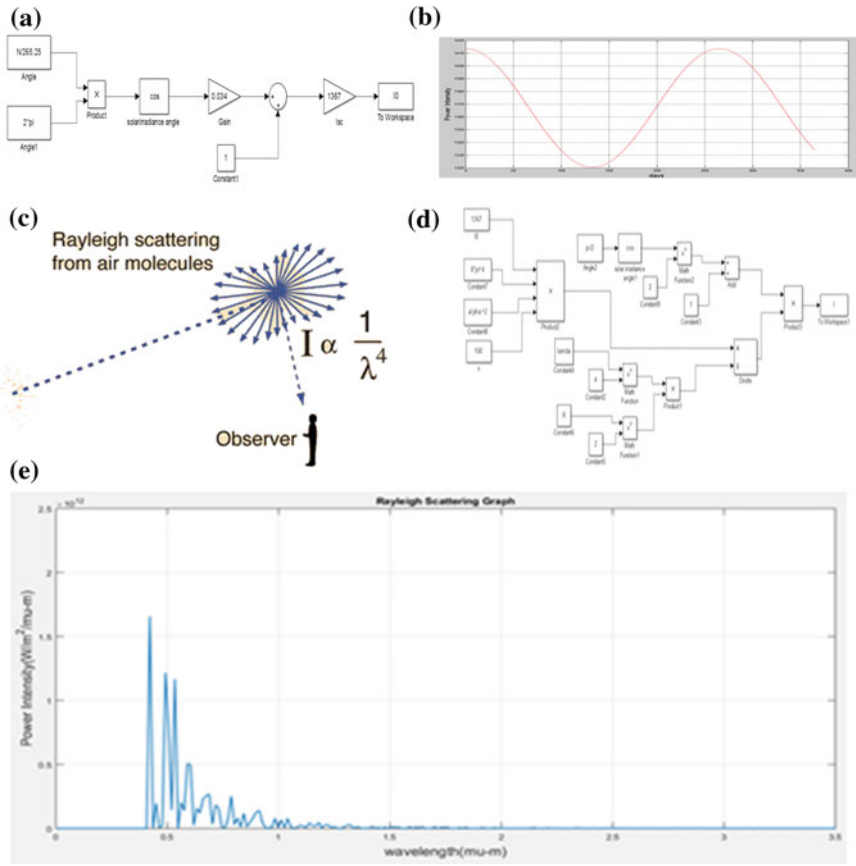


Fig. 1 **a** Simulink Model of power intensity with respect to number of days. **b** Graph between power intensity and number of days. **c** Relation between solar power intensity and sun spectrum. **d** Simulink model of Rayleigh scattering. **e** Graph between power intensity and wavelength

Graph in Fig. 1e is the result from Simulink shown in Fig. 1d showing that power intensity is constant initially at a certain value of wavelength and then at a value 0.5 μm of wavelength power intensity reaches to the maximum value and then decreases again there are certain fluctuations on increasing the values of the wavelength and again power intensity becomes constant at a value of wavelength from 3 to 3.5 μm.

drksujatha23@gmail.com

2.3 Generation Rate of Electron and Hole

$$G = \alpha N0\, e - \alpha x \tag{4}$$

where

N0 Photon flux at the surface (photons/unit area/sec)
α Absorption coefficient;
x Distance into the material.

Figure 2a is showing the simulation of the carrier generation rate of electrons and holes with respect to the above Eq. 4. For a solar cell, Eq. 4 is calibrated to achieve the number of quantity for pairing of hole-electron. For any solar energy extraction unit, the generation will always be maximum on the top surface and the intensity of light will keep on decreasing throughout the material exponentially which is seen through above equation [9].

It is evident that the solar light, which is considered as the incident light, has seven different coloured lights mixed in it with different wavelengths for each. In the context of solar energy, harvesting the amount of energy generated is always different for different wavelengths; this is shown here under [10]. Figure 2b is showing the graph for electron or hole generation rate and showing the decreasing exponentially with the variation in the wavelength of the sun spectrum [11].

$$\alpha = 4\pi K/\lambda \tag{5}$$

Here, materials with higher absorption coefficient more readily absorb photons, which excite electrons into the conduction band [12]. The behaviour of absorbance spectrum is completely opposite to transmittance spectra because as thickness increases absorption decreases due to inverse relation between transmittance and absorption which is shown in above equation [13].
where λ is the wavelength. If λ is in nm, multiply by 107 to get the absorption coefficient in the units of cm^{-1}.

K = Extinction constant.

Figure 3 is showing the graph, which was obtained by the algorithm, which shows that even for those photons, which have an energy above the band gap, the

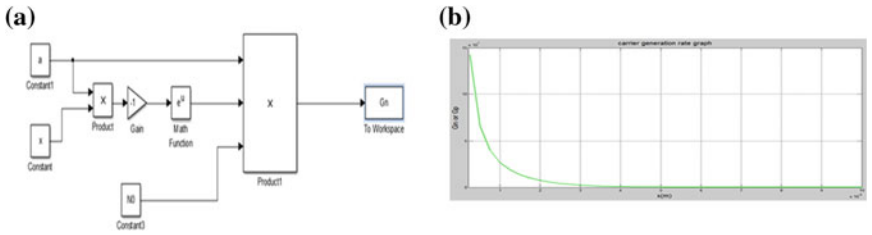


Fig. 2 **a** Simulink model of carrier generation rate of electrons and hole **b** Graph between Gn or Gp and X(m)

drksujatha23@gmail.com

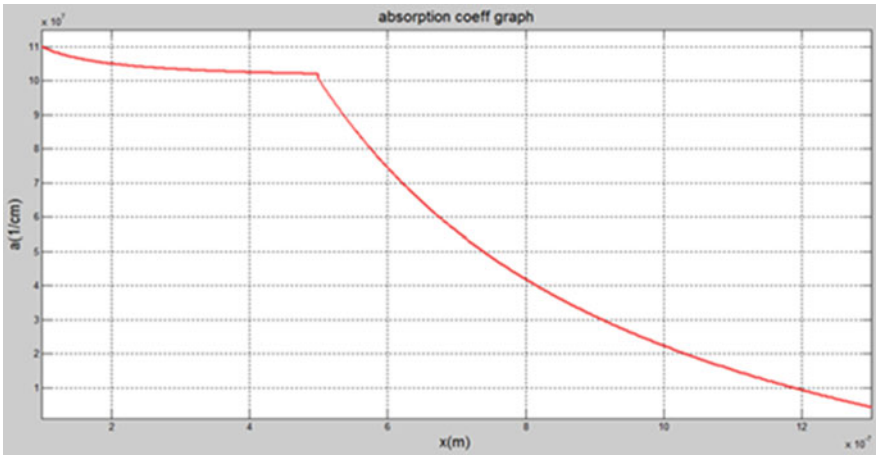


Fig. 3 Graph between wavelength and absorption coefficient angle

absorption coefficient is not constant, but still depends strongly on wavelength. The graphene/GaAs heterostructure system, the drift-diffusion equation and current continuity equation are shown as follows [8, 14, 15]:

$$\mathrm{Jn} = -\mathrm{qn}\,\mu\mathrm{n}\,\partial\mathrm{V} + \mathrm{q}\,\mathrm{Dn}\,\partial\mathrm{n} \tag{6}$$

$$\mathrm{Jp} = -\mathrm{qp}\,\mu\mathrm{p}\,\partial\mathrm{V} - \mathrm{q}\,\mathrm{Dp}\,\partial\mathrm{p} \tag{7}$$

$$1/\mathrm{q}\,\partial\,.\,\mathrm{Jn} - \mathrm{Rn} + \mathrm{Gn} = 0 \tag{8}$$

$$-1/\mathrm{q}\partial\,.\,\mathrm{Jp} - \mathrm{Rp} + \mathrm{Gp} = 0 \tag{9}$$

where Jn, ∂n and Dn are the electron current density, the electron mobility and the electron diffusion coefficient in GaAs, respectively. Jp, n∂p snd Dp are the are the hole current density, the hole mobility and the hole diffusion coefficient in GaAs, respectively. Gn and Gp are electron and hole generation rates due to light illumination, respectively. Rn and Rp are electron and hole recombination rates, respectively. The subsystem shown in Fig. 4a is synthesized for the above Eq. 4 and related by the algebraic expressions for obtaining the resulting graphs.

Here, in Fig. 4b shows the graph elaborating the solar cell's J − V curve. In simulation as visible in Fig. 4a, the doping concentration for charge carriers which are n-type is $1 \times 1022/\mathrm{cm}^3$. The paper has 4.07, 1.42 and 4.8 eV are the values which are kept fixed as affinity level, band gap and Dirac point for GaAs material, respectively. The thickness of Gas and oxide are tGaAs and to x are 350 μm, 70 nm, respectively.

drksujatha23@gmail.com

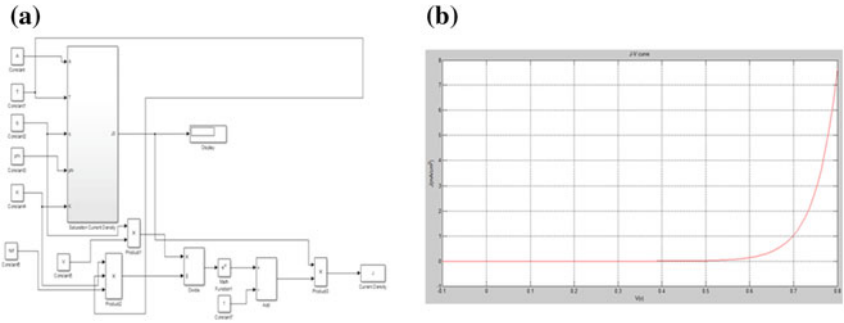


Fig. 4 **a** Simulink model of current density of electron units. **b** Graph between current density (J) and voltage

3 Research Methodology

Here, the research is done on the parameters of the property of GaAs, after synthesizing graphene on the gates of GaAs. To see its efficiency on the working of GaAs gating, all the equations are modulated above and studied by the two most prominent methods, namely Maxwell and Drift-Diffusion equation methodology. For evaluating these equations, the models are being modulated with the help of Finite Difference Method and calibrating their parameters and visualizing it in graphical manner. MATLAB Simulink tool is used to see graphene's effect on GaAs under different conditions to see the properties with respect to different variables.

4 Conclusion

Numerical study on characteristics of graphene' charge carrier for graphene-gated GaAs Schottky junction is done in this paper. It is done on solar cells by using both Poisson's and drift-diffusion equations. Solving them using MATLAB and Simulink Method for solving these equation, we have modulated simulink model and obtained graph by calibrating the parameters. It is shown above that the power absorbed by graphene solar cell is much higher than the power absorbed by the silicon solar cell because the emissivity and collectivity of charge carriers of graphene are much higher than the silicon so by using this we can enhance the efficiency of the Photovoltaic module. Considering sky as a limit, here, there is still lot to be done in PV module using the graphene like studying the output of PV system after installing multilayered graphene structure on it, so that the estimation of output from PV module could be done for installation of layers of graphene. With the help of the ANSYS 2015 software, graphene could be artificially

drksujatha23@gmail.com

synthesized because experimentally, it is expensive and non-economical and difficult to test so can be done with the help of this software the efficiency of the layers could be tested in the future work.

References

1. Song X, Oksanen M, Sillanpää MA, Craighead HG, Parpia JM, Hakonen PJ (2011) Stamp transferred suspended graphene mechanical resonators for radio frequency electrical readout. Low Temperature Laboratory, School of Science, Aalto University, PO Box 15100, FI-00076 Aalto, Finland; Center for Materials Research, Cornell University, Ithaca, New York 14853, USA, vol 12, pp 198–202, 11 Jan 2011
2. Dikin DA et al (2007) Preparation and characterization of graphene oxide paper. Nature 448:457–460
3. Novoselov KS et al (2005) Two dimensional gas of massless Dirac fermions in graphen. Nature 438:197–200
4. Shahicli GG, Warncick I, Ivnri UD, Wu B, Taur Y, Wong C, Chcn CL, Rotlrigucz M, Tang DD, lcnkins KA (1992) High a high performance bicmos techology lisinc; 0.25 pin Cmos and Double Poly Dipolar. In: June 1992, Symposium on VLSI technology digest of technical papers, IEEE confrence, p 2
5. Rosencher E, Vinter B, Luc F, Thibaudeau L, Bois P, Nagle J (1994) Emission and capture 2815 of electrons in multiquantum-well structures. IEEE J Quantum Electron 30(12)
6. Chen W, Li X, Yin W-Y, Lin S, Zhao Z, Li E, Zhou H. Modeling and simulation of graphene-gated Graphene-GaAs Schottky junction field-effect solar cell for its performance enhancement. IEEE Trans Electron Devices 62(11)
7. http://www.itacanet.org/the-sun-as-a-source-of-energy/part-2-solar-energy-reaching-the-earths-surface/
8. http://hyperphysics.phy-astr.gsu.edu/hbase/atmos/blusky.html
9. http://www.pveducation.org/pvcdrom/generation-rate
10. Kageyama T, Kiyota K, Shimizu H, Kawakita Y, Iwai N, Takaki K, Imai S, Funabashi M, Tsukiji N, Kasukawa A (2009) Optical absorption coefficient of carbon-doped gas epitaxial layer by means of propogation- loss measurement of waveguide for long wavelength VCSEL. IPRM '09. In: IEEE international conference on indium phosphide & related materials
11. http://nanotechweb.org/cws/article/tech/38279
12. http://www.pveducation.org/pvcdrom/absorption-coefficient
13. Scharfetter DL, Gummel HK (1969) Large-signal analysis of a silicon read diode oscillator. IEEE Trans Electron Devices 16(1)
14. Li X et al (2015) 18.5% Efficient graphene/GaAs van der Waals heterostructure solar cell. Nano Energy 16:310–319
15. https://www.azonano.com/article.aspx?ArticleID=4565

drksujatha23@gmail.com

Development of IoT for Smart Agriculture a Review



Kamlesh Lakhwani, Hemant Gianey, Niket Agarwal and Shashank Gupta

Abstract The Internet of Things is the hot point in the Internet field. The concepts help to interconnect physical objects equipped with sensing, actuating, computing power and thus lends them the capability to collaborate on a task in unison remaining connected to the Internet termed as the "Internet of things" IoT. With the help of sensor, actuators and embedded microcontrollers the notion of smart object is realized. Wherein these smart objects collect data from the environment of development, process them, and initiate suitable actions. Thus, the Internet of things will bring hitherto unimaginable benefits and helps humans in leading a smart and luxurious life. Because of the potential applications of IoT (Internet of Things), it has turned out to be a prominent subject of scientific research. The importance and the application of these technologies are in sizzling discussion and research, but on the field of agriculture and forestry, it is quite less. Thus, in this paper, applications of IoT on agriculture and forestry has been studied and analyzed, also this paper concisely introduced the technology IoT, agriculture IoT, list of some potential applications domains where IoT is applicable in the agriculture sector, benefits of IoT in agriculture, and presents a review of some literature.

Keywords IoT · Smart agriculture · Agricultural IoT · ITU

K. Lakhwani (✉)
Lovely Professional University, Phagwara, Punjab, India
e-mail: Kamlesh.lakhwani@gmail.com

H. Gianey
Thapar University, Patiala, Punjab, India
e-mail: hgianey@gmail.com

N. Agarwal · S. Gupta
JECRC College, Jaipur, Rajasthan, India
e-mail: agrawal.niket709@gmail.com

S. Gupta
e-mail: shashank96gpt@gmail.com

© Springer Nature Singapore Pte Ltd. 2019

V. S. Rathore et al. (eds.), *Emerging Trends in Expert Applications and Security*,
Advances in Intelligent Systems and Computing 841,
https://doi.org/10.1007/978-981-13-2285-3_50

drksujatha23@gmail.com

1 Introduction of the Internet of Things

ITU (International Telecommunication Union) defined the Internet of things as: "IoT is a technology that mainly resolves the interconnection between human to a thing, thing to thing, and human to human." IoT is a world-shattering technology that signifies the future of computing and information interchange [1]. It is based on the communication between intelligent sensors, RFID (radio-frequency identification), GPS (global positioning systems), infrared sensors, remote sensing, mobile communication, and other communication networks [2]. It refers to a network of objects and is often a self-configurable wireless network [1]. The basic purpose of IoT is to make a huge network by the combination of diverse sensor devices such as GPS, RS, RFID, laser scanner, and networks to comprehend the information sharing of global things. IoT can encompass millions of networked embedded smarts devices also called smart things; these smart things are capable of accumulating information about themselves, their environment, and associated smart devices and interconnect this information to other devices and systems via the all connecting Internet (Fig. 1) [3].

IOT applications include diverse areas including transportation, smart agriculture, atmosphere, marketing, supply chain management, health care, infrastructure monitoring, etc. [1]. To achieve a comprehensive perception, intelligent processing and reliable transmission between information sensing equipment and systems, all physical objects can be individually interconnected and addressable in accordance with agreed protocol according to the needs of different applications [2].

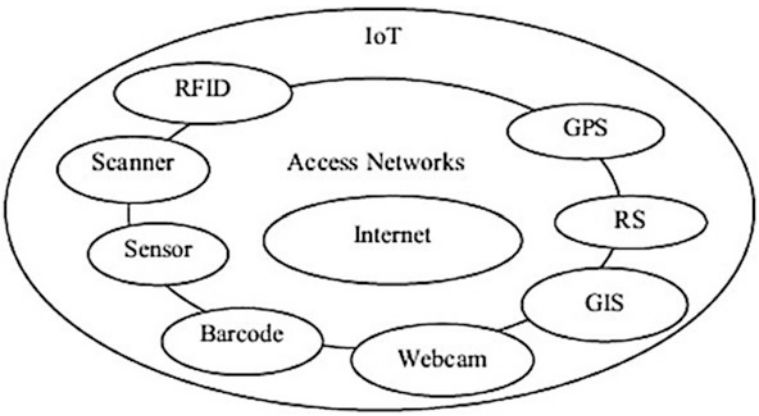


Fig. 1 Conceptual model of IoT [3]

drksujatha23@gmail.com

2 Problem Statements

This paper focused on a basic trade that is Agriculture, which is closely related to the welfare of any nation and the people's livelihood [4]. In India, Agriculture sector is shrinking day by day which disturbs the ecosystem's production capacity. There is a burning requirement to resolve this problem in the area to reestablish vitality and place it back on higher progression.

The reemerging of the worldwide recession has caused flows across both the developed and the developing economies [1]. Agriculture domain required to be more competent and irrepressible to ensure universal food security. Farmers of India are at excessive detriment in terms of technology, size of farms, government policies, trade, etc. The Internet of Things technology can diminish some of the problems of Indian farmers [1].

While in the world, agriculture is experiencing industrialization, it is very significant to develop "agricultural information network". Agricultural information network has become the trend of enlargement for the world's agriculture [5]. In concern of the Indian agriculture development, "agricultural information network" is a major concern in stimulating agricultural development and its transformation.

In India, there are many problems in the agricultural information system. For example, here more importance is given to hardware instead of software and cannot deliver high eminence information to get production requirements of farmers. Besides, information is not adequately used by the farmers of India and the influence of information on a rural area, agriculture, and farmers are not remarkable.

The demand and supply of agricultural products has not been controlled properly, because of the demand and the consumption of the agricultural crops could be anticipated quantitatively, nevertheless, the deviation in crop and production by the weather change, change in cultivated area of farms, damage by insects, disease in crop, etc., could not be truly predicted [5]. To change this situation and endorse the speedy development of agricultural information network, it is required to use the Internet of Things to appreciate smart agriculture [5].

3 Applications of IoT in Agriculture

In the domain of digital Agriculture, IoT supports a variety of applications like soil and plant monitoring, crop growth observing and selection, precision agriculture, irrigation assessment support, greenhouse environment monitoring and control systems, monitoring of food supply chain etc. [1]. Following are the established technologies that are used in applications of IoT in agriculture:

Sensor Technology in Agriculture: Vast variety of sensors are used in agricultural products such as soil moisture sensors, water-level sensors, equipment used to sample the state of the atmosphere at a given time meteorological sensors (monitors

drksujatha23@gmail.com

the current state of atmosphere), heavy metal detection sensors, biosensors (detection of an Analyte), gas sensors (detects presence of gas), and so on.
RFID Technology: RFID is extensively used in animal tracking and identification. It helps to achieve intelligent monitoring, recognizing, traceability of animals, and their management.
Radio Transmission Technology in Agriculture: Self-organizing wireless data transmission can be achieved with ZigBee wireless sensor networks. In large-scale farming, it has been widely used for data transmission.
Intelligent irrigation Technology: Based on satellite positioning network [4] and "shallow wells underground cables + field + automatic irrigation system pipe" [4] technology, it can accumulate irrigation water, irrigation, electricity, and time data to accomplish automation of farmland irrigation and through a complete analysis of information technology software to monitor irrigation.
Technical Quality Safety of Agricultural Products: In the agricultural industrial chain (production–circulation–sales) [4], recording and monitoring of the chain can understand the entire procedure of regulation.
Precision Seeding and Spraying Techniques: Depending on the technology combined with Global Positioning System (GPS) navigation technology, seeding technology, and fertilization at a variable rate, it can achieve identical implementation of the spraying, planting, and refining the consumption of pesticides, seeds, and so on [4].

4 Benefits of IOT in Agriculture

There are various benefits and advantages to use IoT in agricultural sector some of the benefits are as follows:

Efficiency of input: It will improve the efficiency of inputs of agriculture like Soil, Water, Fertilizers, Pesticides, etc.
Cost reduction: It will reduce the cost of production.
Profitability: It will increase the profitability of farmers.
Sustainability: Improves sustainability.
Food safety: It will help to accomplish the Food Safety Mission.
Environment protection: It plays important role in the environment protection.

5 Literature Review

This paper [6] discusses the various applications of IoT and cloud computing in the field of agriculture and forestry. According to the text, the use of IoT plays an important role in smart agriculture. The basic technologies of IoT like laser scanner,

drksujatha23@gmail.com

RFID, photoacoustic electromagnetic sensors, etc. these technologies can be used to make great innovations in agricultural. Basically in agricultural information transmission, precise irrigation, intelligent cultivation control, agricultural product safety, and many more. This paper also focuses some applications of IoT in forestry. IoT can play an important role in forest identification and wood tracking and its management. Finally, this paper concludes that the integration of IoT and cloud computing has become a tendency.

In this research work [2], possible applications of the Internet of Things in agriculture for sustainable rural development has been identified. Various business opportunities related to agriculture domain and its benefits that can be generated, using the Internet of Things is discussed in this text. This literature is intended to stimulus strategy on the acceptance of IoT in agriculture and rural development. According to the literature, developers can use IoT technologies to build country-specific technologies based on the agricultural domain. Development of technology will uplift the standard of people and support poverty alleviation.

In this research work [7], many challenges related to the agricultural domain were, addressed. An architecture was also framed for meeting these challenges. According to the text of this paper, farmers should be guided on the right time during different stages of crop growth. In this research work, a knowledge base is created. This knowledge base has various crop details. These crop details speak about knowledge acquisition, market availability, geospatial data flow and the weather prediction data. Monitoring module includes monitoring of various stages of growing plant, calamity check, planning for irrigation, crop profit calculation, etc. Per day need of water of a plant is calculated using evapotranspiration method. This method is based on the devised algorithm. At last, a comparative study was prepared among several applications existing developed system, having properties like efficiency, the knowledge base, reliability and monitoring modules.

This research work [8] explains the importance of cloud computing in IoT and the importance of these two technologies in Agricultural System. In this paper, it is discussed that IoT is closely correlated to cloud computing. The relation between IoT and cloud computing was explained in such a way that IoT gets influential computing tools with cloud computing. In this research work, an agricultural information cloud is assembled. In this agricultural information cloud, smart agriculture system is constructed through the assemblage of the Internet of Things and RFID. Component of IoT generates a large amount of data like data generated by using RFID, sensors, wireless communication etc. this large amount of data is handled by agricultural information cloud. It is concluded that, in the agricultural information network, hardware resources are integrated into the resource pool for achieving the dynamic distribution of resources and to balance the load, it improves the efficiency of resource use.

In this paper [9], an application prototype for precision farming using a wireless sensor network with an IoT cloud is proposed. In this work, an alert system for the control of water stress of plants using IoT technology was presented. The first part of this paper described the steps of the creation of the decision support system intended for an agricultural community in order to be able to estimate the quantities

drksujatha23@gmail.com

of water required. For irrigation management, the farmer will on the benefit from a dashboard software in the form of a graph, to monitor in real time the variations of the soil conditions and on the other hand, a process of notification by SMS will be transmitted via the application when a critical level is reached to avoid water stress. This application can be improved to make it a very sophisticated one envisages the integration of the method of evapotranspiration to calculate the water requirement of a plant per day in the system of decision support.

In the paper [10], a "Greenhouse monitoring system" with a combination of wireless communications and Internet is proposed. The "greenhouse monitor system", designed using IoT has the definite precision of control and monitor, it is very easy to operate and the interface of this system is user-friendly and this system offers real-time monitoring of the environmental parameters in the greenhouse. This system also has some characteristics like high performance, run reliable and can be improved easily.

This paper [11] explains the architectural components of Internet of Things, shows some application areas where Internet of Things is applicable, discussed about some challenges that have to be discussed along with the securities issues that require consideration like extensive deployment, standardization, interoperability, security of data, efficient spectrum usages and unique identification, gathered object safety, security, and energy consumption. IoT getting rapid momentum due to advances in sensing, actuating, and RFID technologies. It aims at blending the virtual world with the real world seamlessly.

In this research work [12], a platform Phenonet is developed using an open-source platform called OpenIoT. Phenonet is basically a semantically enhanced digital agriculture use case. This paper demonstrated the applications and efficiency of Phenonet in a number of use cases. The researchers demonstrated that how an OpenloT platform can help to handle the challenges encountered by the Phenonet application. In project Phenonet, the basic concept of the collection, validation, processing, annotation, and storing of data captured from smart sensors in the field has been proposed. The related semantic queries, reasoning, and experimental results are presented.

In this paper [13], an application for precision agriculture, a customized architecture for agriculture, based on IoT is presented. This is a cloud-based IoT architecture. This project is applicable to various precision agriculture applications. The research proposed a three-layer architecture. The first layer collects the environmental information and supplies for needed actions. The second layer is a gateway layer, this layer connects the front-end and back-end via Internet or network in which data can be stored and processed. Researchers built a prototype of this architecture to test and illustrate its performance. The efficiency of the proposed architecture is demonstrated by the performance evaluation results.

drksujatha23@gmail.com

6 Conclusion

Issues regarding agriculture, rural area, and farmers have been always deterring India's growth. Agricultural modernization is the only solution to these three problems. Still, India's agriculture is far away from modernization. The use of IoT in agricultural modernization will possibly solve the problems. Based on features of IoT and cloud computing, cloud service, SOA (service-oriented architecture) and visualization technologies can generate huge data involved in agricultural production. RFID with IoT technologies can help to build plant factory that can control agricultural production automatically. A perfect use of modern technology and IoT and blend of them can stimulate the rapid development in the modernization of agricultural system. Use of smart IoT in agriculture could effectually solve the issues concerning farmers, agriculture, and rural area.

According to the above analysis, information technology personnel and agricultural scientist should be encouraged to exchange ideas. Especially, those personals understand planting and understand IT can innovate and promote the modernization of farming. Modernization of farming can improve agricultural production and management, the goal of environmental protection and energy saving could be achieved. By using IoT in agricultural, farmers would be able to understand the current choice of agricultural soil, they would be able to know which crops are appropriate for farming in the current stage, other environmental information of farmland, through intelligent analysis and better management.

In the meantime, the following scenario could be seen: Instead of toiling the field in hot water, farmers would be able to manipulate on computers like a mobile phone or on some intelligent tools, to understand watering, cultivating, seeding, reaping, then they can easily finish heavy farm labor. Continued and rapid development of microelectronic technology, network technology is an opportunity for professionals to actively explore the technological development of modern agriculture. Use of Internet of Things is playing important role in the development of world's modern and smart agriculture which sets a foundation for industrial development.

References

1. Patil VC, Al-Gaadi KA, Biradar DP, Rangaswamy M (2012) Internet of things (Iot) and cloud computing for agriculture: an overview. AgroInformatics Precis Agric (i):292–296
2. Dlodlo N, Kalezhi J (2015) The internet of things in agriculture for sustainable rural development. In: 2015 international conference on Emerging trends in networks and computer commununication (ETNCC), pp 13–18
3. Yan-E D (2011) Design of intelligent agriculture management information system based on IoT. In: Proceedings of the 4th international conference on intelligent computation technology automation ICICTA 2011, vol 1, pp 1045–1049
4. Li J, Gu W, Yuan H (2016) Proceedings of the 5th international conference on electrical engineering and automatic control, vol 367, pp 1217–1224

drksujatha23@gmail.com

5. Lee M, Hwang J, Yoe H (2013) Agricultural production system based on IoT. In: 2013 IEEE 16th international conference computer science engineering, pp 833–837
6. Bo Y, Wang H (2011) The application of cloud computing and the internet of things in agriculture and forestry. In: Proceeding of the 2011 international joint conference on service science IJCSS 2011, pp 168–172
7. Mohanraj I, Ashokumar K, Naren J (2016) Field monitoring and automation using IOT in agriculture domain. Procedia Comput Sci 93:931–939
8. Tongke F (2013) Smart agriculture based on cloud computing and IOT. J Converge Inf Technol 8(2):210–216
9. Karim F, Karim F, Frihida A (2017) Monitoring system using web of things in precision agriculture. Procedia Comput Sci 110:402–409
10. Zhao JC, Zhang JF, Feng Y, Guo, JX (2010) The study and application of the IOT technology in agriculture. In: Proceedings of the 2010 3rd IEEE international conference computer science information technology ICCSIT 2010, vol 2, pp 462–465
11. Patra L, Rao UP (2016) Internet of things-architecture, applications, security and other major challenges. In: 2016 international conference on computing for sustainable global development (INDIACom), pp 1201–1206
12. Jayaraman PP, Palmer D, Zaslavsky A, Georgakopoulos D (2015) Do-it-yourself digital agriculture applications with semantically enhanced IoT platform. In: 2015 IEEE 10th International conference on intelligent sensors, sensor networks information processing ISSNIP 2015, pp 7–9
13. Khattab A, Abdelgawad A, Yelmarthi K (2017) Design and implementation of a cloud-based IoT scheme for precision agriculture. In: Proceedings of the international conference on microelectrons ICM, pp 201–204

drksujatha23@gmail.com

Multi-label Classification Trending Challenges and Approaches



Pooja Pant, A. Sai Sabitha, Tanupriya Choudhury and Prince Dhingra

Abstract Multi-label classification (MLC) has caught the attention of researchers in various domains. MLC is a classification which assigns multiple labels to a single instance. MLC aims to train the classifier for modern applications such as sentiment classification, news classification, and text classification. MLC problem can be solved by either converting into a single-label problem or by extending machine learning methods for solving it. In this paper, the challenges faced during training the classifier which includes label space dimensionality, label drifting, and incomplete labeling are considered for review. This paper also shows the newly emerged data analysis methods for multi-label data.

Keywords Multi-label classification · Active learning · Label drifting Hierarchal MLC

1 Introduction

Supervised learning aims at developing a learning model from a set of instances. Considering X as an instance and L as the label space, the aim of supervised learning is to build a function which can map F (X, L): X → L. Traditional classifications were used with the assumption that each instance is assigned to a single class label from a set of labels (L), L > 1. If |L| = 2, the learning problem is called binary classification problem [1]. Many machine learning algorithms [2] had

P. Pant · A. Sai Sabitha · T. Choudhury (✉) · P. Dhingra
Amity University Uttar Pradesh, Noida, Uttar Pradesh, India
e-mail: tchoudhury@amity.edu; tanupriya1986@gmail.com

P. Pant
e-mail: ashima81k@gmail.com

A. Sai Sabitha
e-mail: assabitha@amity.edu

P. Dhingra
e-mail: princedhingra52@yahoo.com

© Springer Nature Singapore Pte Ltd. 2019

V. S. Rathore et al. (eds.), *Emerging Trends in Expert Applications and Security*,
Advances in Intelligent Systems and Computing 841,
https://doi.org/10.1007/978-981-13-2285-3_51

drksujatha23@gmail.com

been developed to deal with binary classification problem or traditional classification problem. Recently, multi-label classification (MLC) had captured a lot of attention in the research community. In real-world data analysis, there is a need to assign multiple labels to a single label. This approach is called multi-label classification (MLC) and it has the potential requirement in many real-life applications. For example, in a medical field, a patient can have multiple symptoms in a single disease. The news data can be classified under various labels like entertainment, education, and public awareness. MLC problems are quite challenging because of the exponential size of the label space and the dependencies of these labels on each other. Thus, there is a need to understand and address this problem. This paper discusses the various problem transformations and adapted algorithm methods for solving MLC problems. It also discusses the various challenges faced during analyzing multi-label data, the new data analysis approaches emerged from MLC data. There are two types of classification techniques, multi-class classification and multi-label classification. Multi-class classification (MCC) is based on the assumption that every instance could be assigned to only one class label. For example, an employee of an IT can belong to one and only one department; he cannot belong to finance and programming department simultaneously [3]. The multi-label classification aims at assigning a set of labels to every instance. For example, an employee can have marketing, finance, computer, and database skills.

2 Systematic Review

Before preceding with the systematic review, three research questions (RQ) were framed related to multi-label classification. The solutions of the RQ discussed in this paper have been proposed by various researchers. The RQ questions are as follows:

RQ1. What are the various machine learning algorithms/methods that are used for multi-label classification?
RQ2. To identify the challenges faced during classification of multi-label data?
RQ3. What are the trending paradigms for multi-label data classification?

3 Search Strategies for Preliminary Studies

A number of papers related to MLC from various sources were considered for the research work. Boolean OR and Boolean AND strategies were used to search terms with similar meaning and restricting the research.

drksujatha23@gmail.com

Table 1 Research paper collection from different sources

S. no.	Source	No. of result retrieved	No. of relevant papers identified
1	IEEE	19	10
2	Springer	10	5
3	Conferences	20	16
4	Other journals	33	10
5	CiteSeerX	9	3
6	Others	14	10

Information Collection

A total of 105 papers are collected using the abovementioned strategies for review purpose. All the duplicate papers were not considered. Fifty-four papers were found to be relevant and are considered for research. Multi-label classification papers published before 2000 are not considered for this research (Table 1).

4 Result

RQ1: What are the various machine learning algorithms/methods that are used for multi-label classification?

Problem transformation and adapted algorithm are the two methods for solving multi-label classification.

Problem transformation method decomposes the given problem into single-label problems in order to train the classifier [4]. In the adapted algorithm, traditional single-label classification methods are applied to multi-label data [5]. The most common way to perform multi-label classification is by problem transformation method.

4.1 Problem Transformation (PT) Method

In problem transformation method, a multi-label problem is transformed into a single-label problem and the classifier is trained accordingly. The result of the single-label transformation is again transformed to multi-label. There are a number of well-defined multi-label methods like PT1, PT2, etc.

This is explained using an example given below (Refer Table 2).

A. **PT methods**

PT1, PT2, PT3 Method

PT1 is the first problem transformation based on single-label classification. This method randomly chooses a single label of the instance and discards the rest as

drksujatha23@gmail.com

Table 2 Example of multi-label dataset

Instance	L1	L2	L3	L4	L5
1	X	X			
2	X	X	X		
3				X	
4	X	X			X
5		X		X	

Table 3 PT1 and PT2 problem transformation method

Instance	L1	L2	L3	L4	L5		Instance	L1	L2	L3	L4	L5
1	X					⇒	1	X				
2			X				2			X		
3				X			3				X	
4	X						4	DISCARDED				

shown in Table 3. This can lead to loss of important information and affects the accuracy of the result. PT2, the second problem transformation method, uses the result of PT1 and discards the similar instances from the resulted dataset as shown in Table 3.

PT3 problem transformation uses the initial dataset and considers every possible label set combination (Refer Table 4). PT3 takes into consideration the relationship or the dependencies among labels in the label set. It is capable of taking distinct label sets or labels occurred in the label space. PT3 considers frequently occurring, rarely occurring and exceptional label sets with the same preference resulting in unbalancing the single-label classification. Thus, the PT3 approach works slower as compared to PT1 and PT2 (Table 5).

B. Binary Relevance

Binary relevance (BR) is the most researched work on multi-label transformation method. It transforms a multi-label data problem to "L" binary problems where each binary classifier considers each label independently (Refer Table 6).

C. Label Power Set

Label power set eliminates the label independency problem of binary relevance method. It considers each unique combination of labels. This technique uses label ranking to calculate the ranking of the labels [6] (Refer Table 7). Label power set performance can be calculated using the probability of occurrence of the labels.

drksujatha23@gmail.com

Table 4 PT3 problem transformation

Instance	L4	L1 ∩ L2	L2 ∩ L4	L1 ∩ L2 ∩ L3	L1 ∩ L2 ∩ L5
1		X			
2				X	
3	X				
4					X
5			X		

Table 5 PT4 problem transformation

Instance	L1	⌐L1	L2	⌐L2	L3	⌐L3	L4	⌐L4	L5	⌐L5
1	X		X			X		X		X
2	X		X		X			X		X
3		X		X		X	X			X
4	X		X			X		X	X	
5		X	X			X	X			X

Table 6 Binary relevance

Ins	Labels	Label	Label	Label	Label
1	L1	L2	⌐L3	⌐L4	⌐L5
2	L1	L2	L3	⌐L4	⌐L5
3	⌐L1	⌐L2	⌐L3	L4	⌐L5
4	L1	L2	⌐L3	⌐L4	L5
5	⌐L1	L2	⌐L3	L4	⌐L5

Table 7 Label power set

Ins	P(c/x)	L1	L2	L3	L4	L5
L1, L2	0.4	1	1	0	0	0
L1, L2, L3	0.2	1	1	1	0	0
L4	0.1	0	0	0	1	1
L1, L2, L5	0.0	1	1	0	0	1
L2, L4	0.3	0	1	0	1	0
	$\sum P(c/x)L_c$	0.6	0.9	0.2	0.4	0.0

drksujatha23@gmail.com

4.2 Adapted Algorithm

A. Multi-label decision tree

Multi-label decision tree adapts the principle of simple decision tree for classification in the field of genomics using a modified C4.5 algorithm. This requires entropy to split the tree. They recursively use the sum of entropy of the labels for creating a decision tree. Decision tree classification of multi-label instances paved way for the concept of hierarchal multi-label classification.

For a multi-label instance,

$$T = \{(xi,\ li)\ 0 < i \leq n\} \tag{1}$$

The entropy (Ent) of the instance is

$$\text{Ent}\,(T) = \sum p(l)\log(p(l)) + (1 - p(l))\log(1 - p(l)) \tag{2}$$

where p(l) is the probability that instance has a label l and l ε L.

B. TREE-Based Boosting

TREE-based boosting is based on AdaBoost. It aims to reduce the hamming loss by covering and improving the loss function [7]. Weights are assigned to instances after each iteration. The hypothesis is defined by Adaboost.MR. This method is able to detect and remove the outliers, but it is susceptible to noise. AdaBoost-M1, AdaBoost-M2, AdaBoost-MH, and AdaBoost-MR are boosting techniques which can be used for multi-label data.

C. Multi-label K-Neural Network

Multi-label K-neural network (ML k-NN) provides a method to deal with multi-label problem using lazy learning and k-NN. This method follows the concept of error function in backpropagation. K-NN is used for traditional classification or single-label classifications.

The basic concept of k-NN is used for multi-label problems. For a training set "S", let "E" be the training error and "Ei" be the error on (xi, li) and Cij is the actual output on jth label [8].

$$S = \{(xi, Li) | 0 < i \leq m\} \tag{3}$$

$$Ei = \frac{1}{|Li||Li|}\sum \exp\left(-\left(C_k^i - C_1^i\right)\right) \tag{4}$$

$$E = \sum_{l=1}^{M} Ei \tag{5}$$

drksujatha23@gmail.com

D. **Rank SVM**

SVM has been considered the most successful binary classification technique which uses the concept of maximum margin strategy rank [9]. Research algorithms had been proposed for MLC, using popular classification techniques like neural network, evolutionary algorithms like gnetic algorithms.

5 Challenges of Multi-label Data Classification

RQ2. To identify the challenges faced during classification of multi-label data?

With new multi-label learning methods, new challenges are emerging. Many solutions are provided in the literature and are based on assumptions. The accuracy obtained is also based on the nature of the dataset.

A. Dimensionality

As the label space increases, the dimension of the instance space also increases. A dimensionality reduction algorithm aims at removing irrelevant and noisy data [10]. When there exists a high-dimensional label space, dimensionality reduction needs to be applied as a separate data preprocessing step. Dimensionality reduction, for single-label data, is the most intensively researched. It is based on feature extraction and feature selection methods. Dimensionality reduction can be applied to high-dimensional data using a number of unsupervised methods without modifying the data [11], Principal Components Analysis (PCA), Singular Value Decomposition (SVD), Support Vector Machines (SVM), Linear Discriminant Analysis (LDA), Independent Component Analysis (ICA), etc. [11, 12]. PCA is the mostly applied on feature extraction dimensionality reduction method.

B. Data cleaning

There are several issues in data cleaning step which needs to be addressed. For example, in a job recruitment application, data of job seekers are asked to mention their skills. Every seeker enters his skills, irrespective of which domain his skills belongs to. This type of multi-label data requires text analytics, tokenization, stemming, and stop word removal. This can result in loss of important information and degrade the quality of data. Malik and Bhardwaj [13] had provided a way of finding label errors in their research work.

C. Label Dependency

Label dependency or label correlation focuses on the nature of the occurrence and combinations of labels. Many proposed methods only lead to decomposing the multi-label problem to either pair label problem or binary label problem. These approaches had reduced the quality of data [14]. Creating various label combinations may further result in exponentially increased data. Simultaneously, learning from rarely occurring labels becomes a challenge.

drksujatha23@gmail.com

There exist two types of label dependencies, namely, conditional label dependency and marginal label dependency [15]. Conditional dependency tends to capture the dependency of able for a given instance. Marginal dependency is said to exist when product of marginals is not equal to as given in the equation [p(y2) != p (y2|y1) p(y1) p(y2) != p(y1, y2)].

Marginal dependency can be considered as expected dependency. Multi-label classifier processes the multi-label simultaneously, and thereby it affects the performance of the classifier. Hamming loss and subset 0/1 loss are used for calculating the performance matrix. Dembczyński et al. [16] proposed a way of linking label dependencies and loss minimization, in viewing a problem with three different perspectives and minimizing the multi-label loss functions.

D. Label Uncertainty

Uncertainty of labels in real-world applications like recruitment dataset, e.g., the key skills, is to be filled by the users. This results in creation of massive new labels. Every skill can have different names henceforth adding unnecessary new labels. The same label can be repeatedly generated also. In this scenario, the most trending challenge is gaining the knowledge of all the skills of each available domain and using stemming and sub-categorization process for each label.

E. Drifting

As single instance of multi-label data has a number of labels, the interest on labels starts drifting as it is hidden conceptually. For instance, a job seeker changes his job domain or marks a new domain; this is called as "Conceptual drift". Drifting sometimes occurs suddenly and sometimes with a slow rate. Drifting of an instance can be analyzed using instance selection methods. Spyromitros [17] considered these drifting issues as the toughest challenges for any classifier.

F. Data imbalance

It is a very common problem in decision tree and SVM method that some class labels with less frequency have greater importance than frequently occurring labels. According to Min-Ling [18], Charte [19], and Xioufis [20], every multi-label dataset has few labels which tend to be more relevant than most of the commonly occurring labels. This causes label imbalance and has a wide impact on the performance of the classifier. To solve label imbalance problem, methods have been proposed by Wang [21]. His approach could reduce hamming loss but was not able to completely eliminate it. In big data gathering, most of the data are generated by sensors, which are mainly uncertain and imbalance. Extensive research works are being carried out to identify the MLC methods to address these challenges.

RQ3. What are the trending paradigms for multi-label data classification?

Manual labeling of instances is time-consuming. It becomes laborious and impractical when labeling multiple data. Active learning aims at reducing the labeling cost and minimizes the efforts required to label the instances. There is less

drksujatha23@gmail.com

research work on active learning of multi-label data, and limited research articles on multi-label active learning are available.

Mostly active learning is used for labeling single-label data. It considers the label that provides the most valuable information. An active learner system comprises the raw input data and an active learner. These learners may include a classifier and an expert system. The expert system does the analyses for determining the labels [22]. The learner can request supervision of his own choice, and the three active learning methods are query synthesis, pool-based [22, 23], and stream-based method.

In the query synthesis, the learner can query the unlabeled instance and the query generated by the synthesis. In the pool-based active learning method, the learner can evaluate all the instances (unlabelled) before selecting the label. The most informative instance is chosen from the unlabelled pool of instances, which is then converted into a query and sent to the expert. The expert labels the query and sends it to the label training set which in turn is sent to the pool after processing by the learner. In stream-based active learning method [24], the learner decides whether the unlabeled instances should be discarded or sent to the learner for further processing. Most of the methods check all the labels in the label space for an unlabelled instance. This results in very high processing cost. Recently, Rai [24] proposed a framework which selects an instance and pairs it with labels; the expert decides which label instance pair is of the more relevant. AURO-r approaches are able to separate relevant and irrelevant labels, and provide a better label ranking technique as compared to other multi-label active learning techniques.

MIML describes multiple instances and multiple labels linked together. In MIML, instances can be associated with multiple class labels. MIML is the most widely seen in real-life classification, for example, an image classification problem, where an image is segmented based on various instances like semantic instances as shown in Eq. (3). In the past decade, many MIML frameworks have been proposed [25, 26] (Zhang 2010). MIML-SVM converts a single instance into multi-label problem in order to solve it. MIML-boot converts a multi-instance–multi-label problem into multiple instance problems for a single label. MIML-NN framework was built for reducing the loss function. Many multi-instance–multi-label frameworks have been proposed, but they are unable to analyze large volumes of data. Many MIML frameworks have been proposed in order to reduce the loss, but these frameworks are built for either single-instance or single-label problems. MIML-fast has outperformed many MIML problems as it uses linear mapping of labels and label optimization via supervised learning models.

Multi-view multi-label learning:With huge multi-label data available, there has emerged a need of analyzing data from different views. The data for these views can be obtained from different data sources. For example, a human can be identified by fingerprint, iris scan, or lip scan. Many learning algorithms tend to concatenate multi-views of a single instance into a single-instance view. The multi-view learning methods can be classified into three broad categories: co-training, subspace learning, and multiple kernel learning. According to Sun [27], co-training is the first

drksujatha23@gmail.com

proposed concept for multi-view analysis. The method works on three assumptions: (a) Every view is sufficient in classifying and identifying the instance individually. (b) The view is conditionally independent irrespective of the label. (c) Target function of different predicts the same label.

Subspace learning: As the data comes from different sources, the scale of data to be processed becomes large and complex to manage. Subspace learning represents large-scale data in comparatively lower dimensionality such that accurate reconstruction is possible. Subspace learning in multimodal data aims at finding the conditional independence in order to improve the accuracy of the result [28].

Machine kernel learning: Machine Kernel Learning (MKL) is widely used for multi-view problems which require a set of kernel. $K(x_i, x_j)$ is the kernel function used for calculating the similarity between examples x_i and x_j. These kernel functions are used for calculating the dot product in the feature space such that the nonlinear mapping is performed in the input space. Thus, there is a need for multiple kernels. MKL provides multiple machine kernel learning algorithm which enables user to combine different predefined learning methods for each source. Many MKL algorithms have been developed for supervised and unsupervised learning. According to Gonen and Alpaydın [29], all the MKL algorithm for finding the kernel functions falls into five categories: (a) Fixed rules, (b) Heuristic approaches, (c) Optimization approaches, (d) Bayesian approaches, and (e) Boosting approaches.

Hierarchical Multi-label Classification (HMC) and Hierarchy of Multi-label Classifiers (HOMER):

Many multi-label classification methods are now effective in classification of data. Few instances belong to number of classes, and such data needs to be represented in a hierarchal format. Hierarchical multi-label classification is similar to other classification except that an instance can belong to two or more classes simultaneously and that a subclass automatically belongs to the superclass.

Many approaches have been proposed in recent years. Single-Label Classification (SLC) approach or SC approach transforms an HMC into SLC problem by performing classification on every individual class. This results in large creation of data as the hierarchy is not considered and the relevance for classes is ignored. Many learners may have skewed distribution of classes. The second approach is Hierarchal Single-Label classification (HSC) which applies the single-label result in a hierarchal way. In Hierarchical MLC (HMC), the classifier predicts all classes at once. HMC has been widely used for text classification. HOMER uses extended meta-labels learning and balanced distribution of labels. It creates a hierarchal tree using top-down and depth-first approach. The HMC method includes a layer of meta-labels or subset of label space, which leads to increased processing performance.

drksujatha23@gmail.com

6 Conclusion and Future Scope

In this review paper, a number of methods for addressing MLC problems had been discussed. The challenges faced during multi-label data processing are discussed in detail. This paper highlights the new methods which emerged as a need for better analysis in MLC. Many data mining techniques like classification and association are identified that are used to solve these issues. Ensemble techniques like AdaBoost and bagging can be used to solve MLC. The research work needs to understand multi-label data preprocessing for big data analysis, as the classification can become very complicated since the real-world data is incomplete and imbalanced. Data reduction for large dimensional dataset and classifying multi-instance data is also a challenging task.

References

1. Zhang M, Zhou Z (2014) A review on multi-label learning algorithms. IEEE Trans Knowl Data Eng 26(8):1819–1837
2. Chrystal B, Joseph S (2015) Multi-label classification of product reviews using structured SVM. Int J Artif Intell Appl 6:61–68
3. Dekel O, Shamir O (2010) Multiclass-multilabel classification with more classes than examples. In: Proceedings of the 13th international conference on artificial intelligence and statistics (AISTATS), pp 137–144
4. Liu H, Li X, Zhang S (2016) Learning instance correlation functions for multilabel classification. IEEE Trans Cybern 1–12
5. Read J (2010) Scalable multi-label classification. PhD Thesis, University of Waikato
6. Cherman EA, Monard MC, Metz J (2011) Multi-label problem transformation methods: a case study. CLEI Electron J 14(1), Paper 4
7. Amit Y, Dekel O, Singer Y (2007) A boosting algorithm for label covering in multilabel problems. In: Proceedings of the eleventh international conference on artificial intelligence and statistics (AISTATS-07), pp 27–34
8. Zhang M-L, Zhou Z-H (2007) ML-kNN: a lazy learning approach to multi-label learning. Pattern Recogn 40(7), 2038–2048
9. Ahuja Y, Yadav SK (2012) Multiclass classification and support vector machine. Glob J Comput Sci Technol Interdiscip 12(11), Version 1.0
10. Ji S, Ye J (2009) Linear dimensionality reduction for multi-label classification. In: IJCAI'09 Proceedings of the 21st international joint conference on artificial intelligence, pp 1077–1082
11. Sorower M (2010) A literature survey on algorithms for multi-label learning, Citeseerx.ist.psu.edu. http://citeseerx.ist.psu.edu/viewdoc/versions?doi=10.1.1.364.5612 (2016)
12. Varghese N (2012) A survey of dimensionality reduction and classification methods. Int J Comput Sci Eng Surv 3(3):45–54
13. Malik H, Bhardwaj V (2011) Automatic training data cleaning for text classification. In: 2011 IEEE 11th international conference on data mining workshops, pp 442–449
14. Read J, Puurula A, Bifet A (2014) Multi-label classification with meta-labels. In: 2014 IEEE international conference on data mining
15. Dembczynsk J, Waegeman W, Cheng W, Hullermeier E (2010) On label dependence in multi-label classification. In: International Workshop on Learning from Multi-Label Data
16. Dembczyński K, Waegeman W, Cheng W, Hüllermeier E (2012) On label dependence and loss minimization in multi-label classification. Mach Learn 88(1–2):5–45

drksujatha23@gmail.com

17. Spyromitros E (2011) Dealing with concept drift and class imbalance in multi-label stream classification, thesis
18. Min-Ling Z, Li Y-K, Liu X-Y (2015) Towards class-imbalance aware multi-label learning. In: Proceedings of the 24th international joint conference on artificial intelligence (IJCAI'15)
19. Charte F, Rivera A, Jose del Jesus M, Herrera F (2013) A first approach to deal with imbalance in multilabel datasets. In: Hybrid artificial intelligent systems, pp 150–160. Springer
20. Xioufis ES (2011) Dealing with concept drift and class imbalance in multi-label stream classification. PhD thesis, Department of Computer Science, Aristotle University of Thessaloniki
21. Wang H (2016) Towards label imbalance in multi-label classification with many labels. In: Arxiv.org. https://arxiv.org/abs/1604.01304
22. Sheng-Jun H, Chen S, Zhou Z-H (2015) Multi-label active learning: query type matters. In: Proceedings of the twenty-fourth international joint conference on artificial intelligence, pp 946–952
23. Cherman EA, Grigorios T, Monard MC (2016) Active learning algorithms for multi-label data Volume 475 of the series IFIP advances in information and communication technology, pp 267–279
24. Rai P (2016) Active learning, 1st edn., pp 1–24. https://www.cs.utah.edu/∼piyush/teaching/10-11-print.pdf
25. Briggs F, Fern X, Raich R (2012) Rank-loss support instance machines for MIML instance annotation. In: Proceedings of the 18th ACM SIGKDD international conference on knowledge discovery and data mining, pp 534–542
26. Nguyen, C-T, Zhan D-C, Zhou Z-H (2013) Multimodal image annotation with multi-instance multi-label LDA. In: Proceedings of the twenty-third international joint conference on artificial intelligence, pp 1558–156
27. Sun S (2013) A survey on multi-view machine learning. Neural Comput Appl 23(7):2031–2038
28. White M, Yu Y, Zhang X, Schuurmans D (2012) Convex multi-view subspace learning. In: Advances in neural information processing systems (NIPS)
29. Gonen L, Alpaydın E (2011) Multiple kernel learning algorithms. J Mach Learn Res 12:2211–2268
30. Multiple kernel learning (2016) In: En.wikipedia.org. https://en.wikipedia.org/wiki/Multiple_kernel_learning. Accessed Sept 2016

drksujatha23@gmail.com

Virtual Reality as a Marketing Tool



Harry Singh, Chetna Singh and Rana Majumdar

Abstract As of late, virtual reality (VR), as another type of innovation, is creating and inspiring open intrigue. VR innovation can give a counterfeit sensible condition controlled by body developments. It gives intuitive encounters, and it is as yet creating to numerous other new fields like medical and the military. As it can give the simulation of real environment, it can also give motion sickness. It is energizing to investigate that is virtual reality innovation is a better way to view things and how convenient it is for what age group. To check that we conduct an experiment where we pick two different methods to look at an educational institute's campus where we utilize institute's website and Google street view of the institute with virtual reality headset and with the help of a survey and two-way ANOVA test, we try to find out which way is more convenient to what age group.

Keywords Virtual reality · Motion sickness · Google street view
Anova test · Website

1 Introduction

Virtual reality is a produced area that is made with programming and showed to the customer to such an extent that the customer suspends conviction and recognizes it as a bonafide circumstance. On a PC, virtual reality is on a very basic level

H. Singh · C. Singh · R. Majumdar (✉)
Amity School of Engineering and Technology, Amity University Uttar Pradesh, Noida, Uttar Pradesh, India
e-mail: rmajumdar@amity.edu

H. Singh
e-mail: Harrysingh1496@gmail.com

C. Singh
e-mail: cchoudhary@amity.edu

© Springer Nature Singapore Pte Ltd. 2019

V. S. Rathore et al. (eds.), *Emerging Trends in Expert Applications and Security*,
Advances in Intelligent Systems and Computing 841,
https://doi.org/10.1007/978-981-13-2285-3_52

drksujatha23@gmail.com

experienced through two of the five distinguishes: sight and sound. The slightest troublesome sort of virtual the truth is a 3-D picture that can be explored brilliantly at a PC, generally by controlling keys or the mouse with the objective that the substance of the photo advances toward some way or zooms in or out. More complex tries to incorporate such procedures as wrap-around demonstrate screens, honest to goodness rooms extended with wearable PCs, and contraptions that let you feel the show pictures. Many individuals know about the term "virtual reality", yet are uncertain about the employments of this innovation. Gaming is a certain virtual reality application as are virtual universes, yet there are a whole host of occupations for virtual reality—some of which are more trying or irregular than others. There are different ways virtual reality can be used which give immense points of interest to us. These include health care, surgery, military, architecture, art, entertainment, education, business, media, sports, and many more. Recently virtual reality is growing in the field of marketing.

1.1 VR in Marketing

Modernization and revolution are observed as a standout among the greatest means to compact with showcasing. Maximum law agency follows a policy to promote themselves online and used it as video showcasing provisions as a substantial method to drive guests to their sites into realization them. Interestingly, it is judged that up to 96% of B2B establishments have drawn in video promoting, of which 73% recognized a positive ROI affect. The virtual reality ascertains to be beneficial in the video promoting by offering a 3-D dimensional experience. The 360° encounter video showcasing makes the recordings more sensible and clients' spurring. This furthermore provides advertisers with a simpler time while drawing their thoughts. The virtual reality likewise demonstrates another option to video spilling for campaigns, same as the television.

1.2 Google Street View

Google street view is an element of Google Maps that empowers clients to see and explore through 360° flat and 290° vertical all encompassing road level pictures of different urban communities around the globe. The street view highlight can be utilized to take virtual strolls, investigate historic points, or discover shops, eateries, and lodgings. The pictures in street view are acquired from extraordinarily fitted autos that drive through urban areas and urban regions, taking all encompassing 360° accounts of all that they find, including individuals finishing their consistently

drksujatha23@gmail.com

activities. To ensure individuals' protection, Google has actualized innovation that foggy spots individuals' appearances and gives an approach to guests to hail wrong or touchy symbolism for audit and expulsion.

1.3 Motion Sickness Due to VR

Virtual reality torment happens when introduction to a virtual space causes reactions that resemble development illness manifestations. The most broadly perceived signs are general burden, cerebral agony, stomach care, squeamishness, disgorging, pallor, sweating, exhaustion, tiredness, perplexity, and lack of care. Different side effects incorporate postural flimsiness and spewing. Virtual reality affliction is unique in relation to movement infection in that it can be caused by the outwardly instigated impression of self-movement; genuine self-movement is not required. It is additionally unique in relation to test system affliction; non-virtual reality test system disorder has a tendency to be described by oculomotor unsettling influences, though virtual reality ailment has a tendency to be portrayed by confusion.

2 Problem Conceptualization

In the experiment, we will compare the use of virtual reality with normal website. We want to find out which approach attracts people's attention more and how participants feel about both after the experiment. In the end of this research, we will get to know that people of which age group is more likely to use virtual reality as an option instead of normal website.

3 Methodologies Adapted

The methods we will be using are experiment and survey with participants after the experiment and to conclude we will apply two-way ANOVA test to the survey. In the experiment, we will show participants the educational institute's campus with the website and then in virtual reality headset using Google street view. We choose the participants which are looking for a good educational institute, to get the better result to our experiment. We will conduct this experiment with the participants of different age groups, so to get age group as one of the factors in the survey. We did the experiment with exactly 70 participants of each age group that is total of 210 participants and show them exactly the same campus with same VR headset to get the unbiased result.

drksujatha23@gmail.com

We have conducted experiment in the following four steps.

Step1: Participants look at the educational institute's campus in the website.
Step2: Same participants look at the campus through VR headset using Google street view.
Step3: A short survey comes after with some questions prepared.
Step4: We applied two-way ANOVA test to the survey to get the conclusion [1].

Survey questions were prepared and presented as follows:

(1) What is your age?
(2) Out of 10 how much satisfied are you with viewing the campus in website?
(3) Out of 10 how much satisfied are you with viewing the campus in VR headset?
(4) Did you feel motion sickness while using VR headset?
(5) In future which method would you prefer?

3.1 Two-Way ANOVA Test

The two-way ANOVA compares the mean differences between groups that have been split into two independent variables (called factors). The primary purpose of a two-way ANOVA is to understand if there is an interaction between the two independent variables on the dependent variable.

4 Results and Inferences

The result of two-way ANOVA test of the data is given in Table 1. In the result, if we look at the ANOVA table, we can clearly see that there is a significant difference between both the methods that is website and VR, as F value of sample is much higher than its F-critical value. And if we look at summary table, we can see why there is a significant difference, and the average value of virtual reality's score is higher than average value of website's score in all the age groups. The difference is higher in age group (15–30) and (31–45) but there is not much difference in age group (46–60) (Figs. 1 and 2).

The above chart reflects that 17.6% participants feel motion sickness due to VR, whereas 82.4% participants did not feel motion sickness. The chart reflects that 76.5% of the participants will prefer VR method, whereas 23.5% of the participants will prefer website method and 82.4% of the participants did not feel motion sickness, while 17.6% of the participants' feet motion sickness.

drksujatha23@gmail.com

Table 1 ANOVA: two-factor with replication

Summary	15–30	31–45	46–60	Total		
Website						
Count	70	70	70	210		
Sum	420	436	494	1350		
Average	6	6.228571	7.057143	6.428571		
Variance	1.362319	1.135404	1.01118	1.365687		
VR						
Count	70	70	70	210		
Sum	560	536	512	1608		
Average	8	7.657143	7.314286	7.657143		
Variance	0.724638	1.098137	1.030228	1.020643		
Total						
Count	140	140	140			
Sum	980	972	1006			
Average	7	6.942857	7.185714			
Variance	2.043165	1.62261	1.03001			
Anova						
Source of variation	SS	df	MS	F	P-value	F crit
Sample	158.4857	1	158.4857	149.4701	1.47E-29	3.864018
Columns	4.514286	2	2.257143	2.128743	0.120287	3.017515
Interaction	55.25714	2	27.62857	26.05689	2.19E-11	3.017515
Within	438.9714	414	1.060317			
Total	657.2286	419				

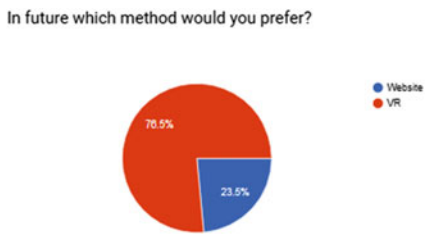


Fig. 1 Graphical representation1

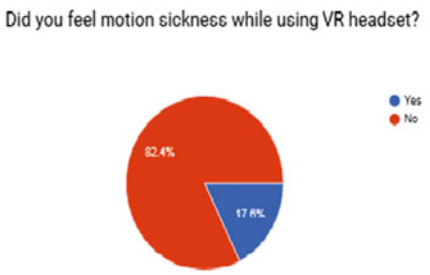


Fig. 2 Graphical representation2

drksujatha23@gmail.com

5 Conclusions

VR helped people to generate more positive response toward the destination. VR is a more engaging form of advertisement/promotion that helps people feels like they are in the destination participating in the activities. Goal advertisers ought to consider utilizing VR to draw in and construct associations with potential guests. Members' interests appear to have been invigorated more in VR than from the website. It appears that virtual reality can make positive recollections and members are more charming in the VR condition. We found that very few members feel motion sickness due to VR and most of the members would like to use VR in future as it was more realistic and interesting. As per the age group, VR generates more positive response by people of age group (15–30) and (31–45). If the marketers want to target the audience of age group 15–45, then they should make use of virtual reality.

Reference

1. Giberson J, Hwan S (Mark) Lee daniel guttentag Maria Kandaurova. Virtual reality and implications for destination marketing

drksujatha23@gmail.com

A Critical and Statistical Analysis of Air Pollution Using Data Analytics



Praveen Kumar, Paras Lalwani, Karan Rathore and Seema Rawat

Abstract Today, in urban areas, people are suffering from many health problems due to air pollution. The health of human beings requires pure air. Today, several diseases are caused due to air pollution. The deaths being seen worldwide are because of air pollution. There are different ways due to which air is being polluted. In Delhi, in order to reduce the air pollution in terms of control measures and pollutant levels, the status of air pollution has undergone through many changes. With increase in air pollution, studies on it have found that there is an increase in all natural cause morbidity and mortality. Increase in industrial activities and vehicular emissions in Delhi are the major cause of air pollution. In the city, the level of pollution of air can be reduced through various ways being identified during the last years. However, to further reduce the air pollution levels, more still needs to be done. The paper analyzes the level of pollution and presents the effects and sources of air pollution in an appropriate manner.

Keywords Morbidity · Eruptions · Deflation · Exposures · Symptoms

1 Introduction

Emitted by both man-made sources and natural sources, the presence of pollutants in the air is responsible for pollution of air today. The organism living on the earth as an example is damaged to a wide extent due to the pollutants present in the air

P. Kumar (✉) · P. Lalwani · K. Rathore · S. Rawat
Amity University Uttar Pradesh, Noida, Uttar Pardesh, India
e-mail: pkumar3@amity.edu

P. Lalwani
e-mail: paraslalwani009@gmail.com

K. Rathore
e-mail: aryanrathore2122@gmail.com

S. Rawat
e-mail: srawat1@amity.edu

© Springer Nature Singapore Pte Ltd. 2019

V. S. Rathore et al. (eds.), *Emerging Trends in Expert Applications and Security*,
Advances in Intelligent Systems and Computing 841,
https://doi.org/10.1007/978-981-13-2285-3_53

drksujatha23@gmail.com

being emitted from vehicles and industries. There are various ways due to which air is being polluted. Motorization, industrialization, and urbanization are the major causes of air pollution. Testing of modern weapons, deforestation, rapid increase of motor vehicles, and installation of petrochemical and chemical plants and thermal plants are some of the other major causes of pollutants being present in the air. One of the biggest causes of air pollution is the greenhouse gases emitted through large-scale industries. Air pollutants refer to the presence of harmful substances in large limits that are commonly gases, liquids, and solids. Sulphur Oxides (SOX), Hydrocarbons (HC), Nitrogen Oxide (NOX), and Carbon Monoxide (CO) are some of the other pollutants that contribute commonly to air pollution. The area being industrialized, the extent of deforestation and the forest cover of the region, the population density of the region, the industrialization extent in the area, the pollutants emitted in the air due to the raw materials used in the industries, the quantum of emissions discharged and in the manufacturing process the engineering technology and science being used, the manner in which the discharges are released into the environment, the manner in which the atmospheric air is mixed with the pollutants, the source of emission of the air pollutants, the topographical and meteorological conditions of the region, the type of air pollutants, and its magnitude are some of the characteristics that influences air pollution of a particular region.

A. Natural Sources
   The sources from nature that pollutes the air include volcanic eruptions, forest fires, deflation of sands and dusts, storms, etc. Some of the natural sources pollutants are volcanic activities, and the atmosphere of earth sometimes gets mixed with land surface pollutants [1]. Some of the examples of pollutants of land are soil particles, sand, salt, etc. Some of the other natural sources of air pollution are comets, cosmic rays, and particles. Some of the other pollutants that cause air pollution in large quantity are vegetation and green plants.
B. Man-made Sources
   Some of the examples of air pollution that are caused through man-made sources are automobiles, industries, domestic sources, power plants, agriculture, etc. Some of the pollutants from man-made sources or human activities are industrial processes; production and manufacturing are some of the pollutants released from industries into air on earth's atmosphere that are the major causes of air pollution. Some of the examples of air pollutants from industries are fumes, smokes, etc. They are emitted in large quantities from large-scale industries. The situation is sometimes also worsened through particulate matters and dust. As a result of activities being done for household purposes, a large amount of harmful substances and chemicals are released into air. Fuelwood, burning of coal gas, and kitchen gases are some of the pollutants released into air through domestic sources. The pollutants that are released from vehicular sources such as various automobiles get mixed with atmospheric air and causes air pollution. Fumes, gases, emissions from surface, and smoke are some of the examples of air pollution caused through vehicular sources. Some of the pollutants are also released into air through the activities that are done for

drksujatha23@gmail.com

agriculture purposes. On the fields of agriculture when certain chemicals are sprayed such as herbicides, pesticides, and insecticides, these also cause air pollution after they get mixed with atmospheric air. The environment of earth is severely damaged because of power plants that are based on fossil fuels. Fossil fuels are one on which the people are heavily dependent. The earth's environment is negatively affected because power plants generate large amount of heat [2]. The atmospheric air gets mixed with fly ash that is being discharged through burning of fossil fuels, and this causes air pollution.

2 Literature Review

Population growth in urban cities is related to power, distribution, health, education, and pollution, pollution in specific. Air pollution results in various health hazards and it also leads to global warming. As shown in table below, it has been identified that the pollution of air is affected by four pregnancy outcomes invested through twelve epidemiologists. IUGR (6 studies), PTD (4 studies), VLBW (2 studies), and LBW (4 studies) were the outcomes being assessed. Methodological features and similar designs were shared among many studies. One ecologic study, one case–cohort study, one case–control study, and nine population-based cross-sectional studies were being identified by twelve investigations being done. Vital records ascertain confounders and data on outcome. The studies work out to collect one ascertained information and desired information through medical and records and personal interviews. In the analysis, all studies were confounded through proper adjustments. The data sources influence the confounding and confounding being addressed varied among different studies. Very often the confounding being assessed were found to contain adjustments for infant and maternal characteristics. Weight gain, alcohol use, and smoking were some of the potential confounders that had been addressed by studies through other sources of data. In geographically defined regions, the pollutant's concentrations were estimated through assessment of the exposure by studies being done [3]. The exception was the sole case–control study, which instead employed environmental transport modeling to estimate exposures for the home where the mother resided at the time of the birth of her infant. When evaluation was done, SO_2 and particulate matter were found to be the ambient contaminants. Related to PM10 standard, the hypertension prevalence against 9.5% in controls was found to be 36% in Delhi. Skin irritation, eye irritation, and chronic headache had significantly higher levels in Delhi. A study on air pollution has found that with increase in air pollution in Delhi, there is increase in mortality. Even when another study was conducted, in association with respiratory problems, gaseous pollutants were found to be at higher level [4].

drksujatha23@gmail.com

3 Methodology

Pollution of air can be controlled through some of the effective ways that are as follows:

Methods of Source Correction: Pollution of air is mainly caused through industries. As an example at the source itself, control of pollution of air can be accomplished by choosing such methods that minimize the potential of air pollution. Equipment for pollution control: There are chances that controlling the atmospheric pollutants may sometimes be not able to control the pollution of air at source itself. In those cases, the removal of main gas stream from pollutants that are gaseous is essential. This can be performed through pollution control equipment being installed. At high concentration, these pollutants are present at the source itself and from the source as the distance of these pollutants increases, through diffusion with atmospheric air, causes the dilution of these pollutants [5]. Control equipment of pollution can be classified as particulate contaminants control devices and gaseous contaminants control devices (Fig. 1).

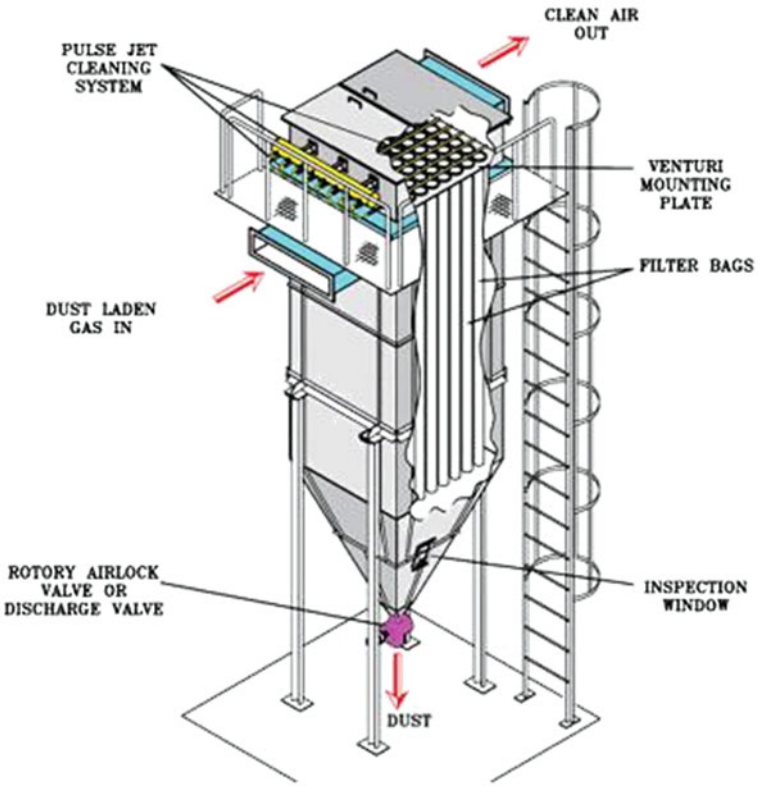


Fig. 1 Air pollution control system

drksujatha23@gmail.com

3.1 Air Pollutant's Diffusion

The atmospheric pollution can be controlled through another approach that is the dilution of the contaminants. If the source of pollution releases contaminants in large quantity, then this results in air pollution. However, if the contaminants are released into atmosphere in small quantity, in that case, pollution of air is unnoticeable. However, the usage of tall stacks that disperse the contaminants and penetrate the upper atmospheric air can be highly helpful in reducing the ground-level pollution and thereby the atmospheric contaminants can be easily diluted. The height of nearby structures should be usually 2–21/2 times more than the height of stacks. Directions of wind, speed, and atmospheric temperature are the various factors that influence the pollutants dilution. The method brings about long-range effects in short-term contact that are highly undesirable. This is one of the disadvantages of this method. The reason is that there is possibility of contaminants to come down at some distance from source. The contaminants are not completely diluted. They are diluted only to certain level.

4 Data Analysis on Air Pollution

4.1 Statistical Analysis of Air Pollution

In 1997, the pollution of air has been found to be one of the major areas of concern by Ministry of Forests and Environment. The study reviewed that there were deteriorating conditions in Delhi due to air pollution. The emission of air pollutants in Delhi was found to be 3000 metric tons every day. Consequent upon the rise in vehicular pollution, in 1996, there was an increase of 92% with regard to the concentration of carbon monoxide. About 12% contribution of industrial units along with other air pollutants is because small-scale industries contribute the highest cluster in Delhi. More than 3.4 million is the vehicular pollution being estimated reaching 7% growth rate.

The air quality is generally measured using PM10 standard. The particles having diameter of 10 Âμm or less are included in PM10 standard. Because of the ability of these particles to reach the respiratory tract and that too up to lower regions, they are responsible for adverse health effects. Premature death, cancer, damage to lung tissue, and effects on respiratory and breathing systems are some of the major concerns included from exposure to PM10. The particulate matter mainly affects person who are old, children, and others. By almost 10 times at 198 micrograms per meter cube after Kanpur and Ludhiana, the Delhi exceeds as reported by the study done by World Health Organization in 2011. In Delhi, air pollution both indoor and outdoor was mainly because of industrial activities and vehicular emissions.

drksujatha23@gmail.com

5 Implementation

5.1 Data Analysis of Various Areas of Delhi

1. **Janakpuri**

See Fig. 2.

According to study, there is a continuous rise in pollutants concentration mainly particulate matter. It has been observed that there are various sources of pollution for PM2.5 and PM10. For increasing concentration of pollutants, winters are responsible as they are dominated by low wind conditions with ground-based inversion, cold and dry air. Since warm air layer acts as a lid on top of earth's surface, close to earth's surface high concentration of pollutants is trapped in. Furthermore, during January and December, meteorological conditions and vehicular pollution have been found in Delhi due to dense smog formation.

2. **Nizzamudin**

See Fig. 3.

Generated by breathe air quality monitoring devices of IndiaSpend and according to particulate matter's analysis, there is a rise in pollution of air in Delhi by 50% in the beginning of 2016 as compared to the last week of 2015 based on odd–even measure. Odd–even measure policy on Delhi's air needed to be understood properly as an increase is seen in the level of PM2.5. In order to make the air clean, some of the additional measures suggested by IndiaSpeed were closure or upgradation of obsolete power plants.



Fig. 2 Air quality in Janakpuri

drksujatha23@gmail.com



Fig. 3 Air quality in Nizamuddin

3. **Sarojini Nagar**

See Fig. 4.

It can be said that Delhi is highly polluted based on above observations in terms of particulates. It has been estimated that 72% of the total air pollution is due to vehicles. Due to changes in general wind direction, high wind velocities, and precipitation, the levels of respirable suspended particulate matter are found to be lower in months of monsoon. In winter months, there is a higher level of all the air pollutants. As compared to other seasons, during the winter months, the concentration of pollutants is higher since the pollutants are not widely dispersed. The reason for this is that atmospheric [6] dispersion is minimum and average mixing



Fig. 4 Air quality in Sarojini Nagar

drksujatha23@gmail.com

height is lower. Followed by several one of the most polluted cities in the country is Delhi. Concentrations of PM10 have been found to be increasing [7, 8]. Through RTI, the data of PM10 has to be sourced out as been observed. On State pollution control boards or Central Pollution Control Board (CBCP), there is no common format in which the data for all cities are available as said by Sunil Dahiya, campaigner, Greenpeace India.

6 Result and Analysis

(1) Nitrogen Dioxide
Photochemical smog is mainly caused by this reddish-brown irritating gas. Nitrogen dioxide gets converted into nitric acid that is HNO_3 in the atmosphere. Power plants and smoke due to burning of fossil fuels generally release this pollutant into air. Fabrics can get damage due to this pollutant, and it can also harm human beings by causing irritation to lungs. NO_2 can also damage ancient documents, soil, trees, and lakes.

(2) Carbon Monoxide: The fossil fuels that are incompletely combusted release carbon monoxide into atmosphere. Also smoking cigarette releases this pollutant into air. The pollutant is a highly poisonous gas that is odorless and colorless. The oxygen-carrying ability of red blood cells to tissues and other cells of body get reduced due to this air pollutant as a result of which anemia and headache can affect the health of person. Irreversible brain damage, coma, and even death can be caused if the pollutant is released at high levels into the atmosphere [9].

(3) Sulphur Dioxide: Coal and oil generally contain sulphur. When such fossil fuels are combusted, they release sulphur dioxide into air. Acid deposition involves this pollutant as a major component because sulphur dioxide gets converted into sulphuric acid in the atmosphere. Due to the release of this pollutant, healthy people may suffer from breathing [10, 11] problems. The pollutant also affects the environment because aquatic life is highly affected and ancient documents are deteriorated due to acid deposition. Their visibility too gets reduced.

(4) Suspended Particulate Matter: For short to long periods, a variety of droplets commonly known as aerosols and particles that are released into atmosphere are included as suspended particulate matter. Burning of other fuels and diesel in construction, agriculture, unpaved roads, and vehicles, coal burning in power and industrial units are some of the human sources for suspended particulate matter. Bronchitis, asthma, cancer, and throat irritation are some of the health effects of suspended particulate matter. Acid depositions that can harm the aquatic life and deteriorate the ancient monuments are some of the environmental effects of the SPM. Suspended particulate matter also reduces visibility.

drksujatha23@gmail.com

7 Conclusion and Future Scope

Air pollution has been found to be one of the major curses on the health of public. To control the level of air pollution, various steps had been taken during the last 10 years. The participation of community along with the government's efforts is necessary to control the pollution of air. At traffic intersections, people must switch off their vehicles. If job opportunities get developed in suburban and peripheral areas, the air pollution will get reduced as this will reduce the migrants to Delhi. Today, health is important to every citizen of India. If health is proper, then only you can gain wealth. This is an important saying that everyone knows. The analysis being conducted focuses on several steps that work efficiently to reduce the air pollution.

References

1. Graff Zivin JS, Neidell MJ (2011) The impact of pollution on worker productivity. National Bureau of Economic Research, April 2011
2. Gemmer M, Xiao B (2013) Air quality legislation and standards in the European union: background, status and public participation. Adv Clim Change Res 4:50–59
3. Romley JA, Hackbarth A, Goldman DP (2010) Impact of air quality on hospital spending. Dana P. Rand Corporation
4. Volk HE, Lurmann F, Penfold B, Hertz-Picciotto I, McConnell R (2013) Traffic related air pollution, particulate matter and autism. JAMA Psychiatry, January 2013
5. Moreno E, Sagnotti L, Dinarès-Turell J, Winkler A, Cascella A (2003) Biomonitoring of traffic air pollution in Rome using magnetic properties of tree leaves. Atmos Env 37: 2967–2977
6. Panko JM, Chu J, Kreider ML, Unice KM (2013) Measurement of airborne concentrations of tire and road wear particles in urban and rural areas of France, Japan and the United States. Atmo Env 72
7. Caselles J, Colliga C, Zornozo P (2002) Evaluation of trace elements pollution from vehicle emissions in pentunia plants. Water Air Soil Pollut 136:1–9
8. Jaswal K, Kumar P, Rawat S (2015) Design and development of a prototype application for intrusion detection using data mining. In: 2015 4th international conference on reliability, infocom technologies and optimization (ICRITO) (trends and future directions), pp 1–6. IEEE
9. Zawar-Reza P, Kingham S, Pearce J (2005) Evaluation of a year-long dispersion modelling of Pm10 using the musicale model TAPM for Christchurch, New Zealand. Sci Total Env 349:249–259
10. Agrawal M, Singh J (2000) Impact of coal power plant emission on the foliar elemental concentrations in plants in a low rainfall tropical region. Env Monit Assess 60:261–282
11. Barman SC, Kumar N, Singh R (2010) Assessment of urban air pollution and its probable health impact. J Env Biol 31(6):913–920

drksujatha23@gmail.com

Conceptual Structure of ASMAN Framework to Compare SaaS Providers



Mamta Dadhich and Vijay Singh Rathore

Abstract Software as a service has been catching the attention of SaaS consumer in recent years and it appears to be an essential need of SMBs/individual. In this paper, we proposed the ASMAN (Appropriate Selection of SaaS Model According to Need) framework to select suitable SaaS service. We recognized the required phases for ASMAN Framework and describe three-layered architecture with their significant role in the selection process. We designed a conceptual structure to explain the working of ASMAN and discussed the overview of ASP.NET MVC technology with the reference of ASMAN Framework based on some essential parameters, named; Availability, cost, reliability, speed, and usability (ACRSU). These parameters are the core attributes of this framework to evaluate the service providers.

1 Introduction

In this chapter, we had recognized the quality of model and requirements of the proposed model for selecting suitable SaaS, that is a popular phenomenon in cloud computing and also consent to the identification of several mandatory parameters and an algorithm to compare the attributes. We will discuss the proposed design of a "Framework named ASMAN (Appropriate Selection of SaaS Model According to Need) to select suitable SaaS" for software as a service system from the viewpoint of the SaaS consumer.

M. Dadhich (✉)
Department of Computer Science, IIS University, Jaipur, Rajasthan, India
e-mail: mamtadadhich76@gmail.com

V. S. Rathore
CSE, Jaipur Engineering College & Research Centre (JECRC), Jaipur, Rajasthan, India
e-mail: dr.vijaysinghrathore.cse@jecrc.ac.in

© Springer Nature Singapore Pte Ltd. 2019

V. S. Rathore et al. (eds.), *Emerging Trends in Expert Applications and Security*,
Advances in Intelligent Systems and Computing 841,
https://doi.org/10.1007/978-981-13-2285-3_54

drksujatha23@gmail.com

2 Literature Review

The literature review enables a researcher to study methodologically analyze and synthesize quality literature, providing a firm foundation to the research topic and the selection of research methodology, and demonstrating that the proposed research contributes something new to the overall body of knowledge or advances the research field's knowledge base (Levy and Ellis 2006).

Quality Metrics developed to evaluate SaaS key attributes define metrics for each quality attribute and provide a description including formula, value range, and relevant interpretations. Analytic Hierarchy Process (AHP) technique for prioritizing the product features and also for expert-led scoring of the products (Godse and Mulik [1]. generated ROSP Algorithm find the optimal fuzzy value for each cloud service provider and allot the user, CSP with the maximum fuzzy value. This job scheduling algorithm makes possible the cloud middleware (broker) to decide capability of CSP by using Rough Set analysis [2].

A comparative overview of existing research work is presenting as below.

	Author description	Existing methodology	Qualities of existing models	Unfavorable features of existing models
1.	Godse and Mulik [1]	Analytical hierarchy process (AHP) weight calculating approach for each attribute and develop three level hierarchies for comparison matrices to judge weight	1. It allows judgment separately on each of several properties 2. Comparison methodology implies three different parts respectively; parameters comparison, product comparison and combines the results acquired from first two parts to rank the products	1. Wide range of selection parameters includes technical and non-technical attributes 2. A little bit doubt regarding the result due to the ratio scale, get after calculating weight, because this pairwise comparison depend upon measurement scales used and the guidelines with reference to the comparison process
2.	Lee [3]	Quality model determines quality matrices for each attribute, and provides a description including formula, value range, and relevant interpretations	1. Evaluate usefulness and practicability of the five metrics by applying correlation, Consistency and Discriminative power to evaluate highest Efficient SaaS 2. A strong linear association between quality attributes and their metrics, service providers analyze their services and may predict their ROI (Return on investment)	1. This quality model more beneficial for SaaS provider instead of SaaS consumer 2. This comprehensive model is not fit for SaaS user because quality matrices evaluate a general comparison of SaaS key attributes and supports SaaS providers 3. Cloud service provider prefers quality model

(continued)

drksujatha23@gmail.com

(continued)

	Author description	Existing methodology	Qualities of existing models	Unfavorable features of existing models
3.	Mahrishi [2]	Rough set model by introducing a middleware (cloud broker) find out potential of CSP by using Rough Set analysis	1. Cloud middleware approach, introduced to meet SaaS services where the cloud brokers are responsible for the service selection 2. This model implies on two different part of evaluation, i.e., Part 1 of the algorithm involved to extract value of providers and second part is for executing at client side and extract the optimal CSP on the basis of achieved fuzzy values	1. The parameters which are considered as standard parameters are not essential and mandatory parameters, these are; Data Operation, Risk management, Legal Issues, Compliance and Audit, Inter-portability, and Portability 2. Complexity will increase of an already complex system and brokers are founded to address security and compliance therefore not suitable for security point of view. By introducing a middleware, CSB
4.	Amrutha and Madhu [4]	Rank Cloud Framework Model; Proposed brokering method implies on broker algorithm, selects some QOS parameters and proposed matrices for ranking cloud	1. The proposed architecture uses response time, suitability, interoperability, and cost of services for ranking CSP's 2. Middleware approach, i.e., cloud broker between provider and end user	1. A generalized framework for cloud computing. Therefore, it is the broader area of incorporate to this model 2. The problem reaming same with cloud broker as we discussed above 3. There is no specific category of cloud was considered while adopting best one so broader area to implement this method
5.	Boussoualim (2015)	The MCDM approach, calculating weight of attributes in three levels; hierarchy is developed by applying AHP technique Algorithm introduced to the device for a middleware; cloud broker	1. It is again a mathematical model based on a pairwise comparison of parameters by calculating weight of given attributes 2. Line-of-business services offered which mainly focused on the applications are; Content Management (CM), office software, Customer Relationship	1. This methodology basically is more suitable as non-technical criteria of selection parameters 2. Cloud Broker was proposed in this method to meet user's requirement so it is another cause of system complexity and security problem 3. There is more concern about non-technical

(continued)

drksujatha23@gmail.com

(continued)

	Author description	Existing methodology	Qualities of existing models	Unfavorable features of existing models
			Management (CRM), Business Intelligence (BI), Enterprise Resource Planning (ERP) and Supply Chain Management (SCM), based on two criteria their functionality and preferences	issues and security problem because of middleware.
6.	Jagli et al. [5]	Proposed Quality model; It is a method to exhibit an algorithm, for a particular decision	1. A decision tree demonstrates the influence of attributes with each other from a given set of attributes and the working process 2. The sample data set is created based on user feedback 3. Considering customizability as major parameter to influence other parameters	1. Dependency upon user's feedback 2. More concern about customizability as key attribute to meet higher maturity model 3. Customizability is explored to show how it is associated with other key attributes

3 General SaaS Adoption Criteria

The selection process is generally based on the concert of the multicomponent facet of different services, therefore, the multicriteria decision-making (MCDM) problem arises, the client has to determine the different alternative level for every performance facet, therefore the selection result can return the user's customized requirement and deliver the functionality of application to the end user.

Figure 1 shows the basic approach that can make use of SaaS application. The basic components of this architecture are end user, SaaS service providers with a list of SaaS applications and these services are hosted by provider at the user's access device via web browser. The architecture is followed by a general three-tier architecture, respectively, which has three layers: end user (web browser), application layer (APIs: application program interface), and data layer (storage of applications and services).

Three-layered architecture consists of the following components to host a SaaS service to the end user:

3.1 Service Consumer Level (End User)
3.2 Middleware Level (APIs; Application Programming Interface)
3.3 Service Provider Level (List of SaaS Services)

drksujatha23@gmail.com

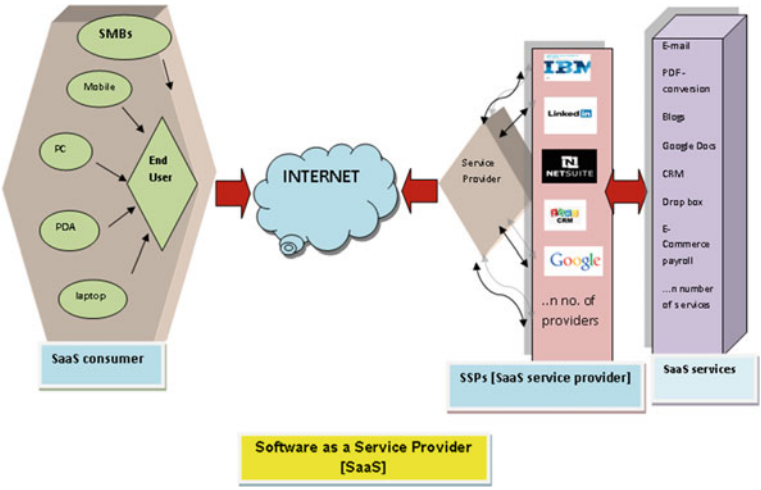


Fig. 1 Three-layered SaaS architecture

3.1 Service Consumer Level (End User)

The main component of this architecture is the end user who can access application as per need through a well-known tool web browser and the whole refined application is available on-request from some SaaS provider. The application is presented in the SaaS cloud data center (not in consumer's ground datacenter) and can be accessed from each browser.

3.2 Middleware Level (APIs; Application Programming Interface)

Middleware technology establishes a link between end user and host. Application Programming Interfaces (APIs) is one of the most important aspects essential to adopt software in cloud computing remote based architecture for their successful use of application across the infrastructure with needful use of services.

3.3 Service Provider Level (List of SaaS Services)

The SaaS service provider does establish a link directly for business with the end user. Vendor is answerable for any kind of requirement need to build an infrastructure such as firewall, load balancer, compute power, storage, etc.

drksujatha23@gmail.com

4 ASMAN Conceptual Structural Model

ASMAN conceptual model is the fundamental approach that we proposed for SaaS selection, in this approach SaaS consumers who may not have sufficient knowledge of how the SaaS services are scheduled according to user's requirement (a user can go with any kind of parameter after comparison as per his/her priority like speed or cost) and how the services are deployed and provided to a user. It holds opposing view from the conventional SaaS selection process. Particularly, the SaaS group of service analyzation because it focused before a general cloud service criteria including PaaS and IaaS. In this section, we will discuss about the conceptual view of SaaS selection structural design to understand the working of ASMAN model which is based on the ASP.NET Model View Controller technology (Fig. 2).

SaaS consumer generates HTTP request by adding required SSPs (not more than three) and their and required service to compare them. MVC architecture has three main components to implement the proposed algorithm namely, model, view, and controller where models deal with data of ASMAN architecture such as list of SSPs, list of parameters and list of services. These are independent data model because there is not a connection among other component of framework.

Controller is the component to operate the different actions over user requests and the data like searching, filtering, and comparing in format of the model object

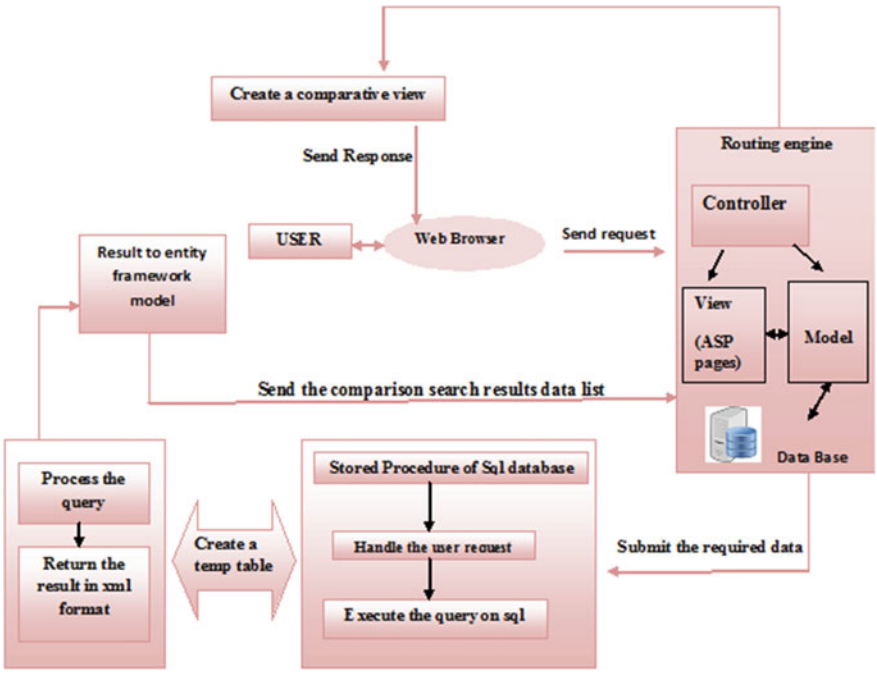


Fig. 2 ASMAN—conceptual structural model

drksujatha23@gmail.com

which are combined with views and send as response to the client, for instance, modals based on types like SSPs, ACRSU essential parameters, and SaaS services, which are pulled out as per request nature and sent with views as response to the client.

View is the user interface by which a consumer can generate a request and get the response as per their requirement.

Comparison search result serves to the end user, which includes the list of SSPs along with their comparison charts of five essential parameters after applying SCAA algorithm.

5 Conclusion

However, there exist some techniques to select appropriate cloud service provider. At the same time, there is a strong need to select suitable SaaS provider to fulfill user's needs by evaluating some quality of parameters [QoP]. These parameters are mandatory to evaluate together within a cloud environment. The aim of research is to facilitate SaaS consumer to compare the available SaaS provider and get appropriate SaaS service from them. ASP.net platform will be used to implement this SaaS comparing algorithm by using code first approach. This research has three major components respectively SaaS essential parameters, SaaS scheduling algorithm to compare suitable SaaS providers among many.

References

1. Godse M, Mulik S (2009) An approach for selecting software as a service (SaaS) product. In: IEEE international conference on cloud computing, pp 155–158
2. Mahrishi M (2013) A novel approach for selection of appropriate software as a service in cloud computing. J Ind Eng Manag 70:48–51
3. Lee JY, Lee JW, Cheun DW, Kim SD (2009) A quality model for evaluating software-as-a-service. In: Cloud Computing Seventh ACIS International Conference on Software Engineering Research, pp. 261–266
4. Amrutha KK, Madhu B, Tech RM (2014) An efficient approach to find best cloud provider using broker. J Adv Res Comput Sci Softw Eng Res Pap 4(7)
5. Jagli D, Purohit S, Chandra NS (2016) Evaluating service customizability of SaaS on the cloud computing environment. Int J Comput Appl (0975 – 8887) 141(9)
6. Whaiduzzaman Md, Gani A, Anuar NB, Shiraz M, Haque MN, Haque IT (2014) Cloud service selection using multicriteria decision analysis. Sci World J Art ID 459375

drksujatha23@gmail.com

A Case Study of Feedback as Website Design Issue



Jatinder Manhas, Amit Sharma, Shallu Kotwal and Viverdhana Sharma

Abstract In today's scenario, the websites play a very important and prime role in communicating organizational policies and vision to the entire globe. Rigorous and extreme efforts are required from different organizations to concentrate more on design part of the website to make them more beautiful and informative in nature. Websites are acting as online agent these days to facilitate the user groups with different types of activities without making them physically visiting the concerned organization. Webmasters plays pivotal role in designing websites and they are required to design websites after properly studying the user behavior. Different standards/guidelines have been recommended by various organizations to help webmaster in designing user-centric website compatible with all latest trends and technologies. Feedback facility on a website helps users to convey their grievances to the organization and at the same time helping organization in understanding user behavior in more close way. Authors have developed an online tool by using .NET Framework after thoroughly investigating different recommended guidelines to study feedback facility as Design issue while designing different kinds of websites. To study this, five different kinds of websites were undertook that includes Government, Commercial, Educational, Social networking, and Job portals. The automated tool developed by author function on the basis of W3C guidelines that are prescribed in document WAI 1.0 (Panta, Web design, development and security. Youngstown State University, 2009) [4]. The automated tool extracts the complete website code and then supplies that code to the parser for the purpose of rendering it thus by producing a result to determine the presence and absence of feedback

J. Manhas · A. Sharma · S. Kotwal (✉)
University of Jammu, Jammu, India
e-mail: shallukotwal25@gmail.com

J. Manhas
e-mail: manhas.jatinder@gmail.com

A. Sharma
e-mail: amitsharma.ju@gmail.com

V. Sharma
Mody University, Jaipur, India
e-mail: viverdhana.17@gmail.com

© Springer Nature Singapore Pte Ltd. 2019

V. S. Rathore et al. (eds.), *Emerging Trends in Expert Applications and Security*,
Advances in Intelligent Systems and Computing 841,
https://doi.org/10.1007/978-981-13-2285-3_55

drksujatha23@gmail.com

facility in a given website. Results show that out of the five different websites undertaken for the study, the government websites shows the maximum positive results whereas other categories show the mild presences of feedback facility on their websites.

Keywords Website · Design · Feedback · W3C · .NET · C#

1 Introduction

Websites are nothing but an effective source of interaction with the organization by different types of users worldwide. Websites act as an online agent through which users can avail all types of facilities and can get there work done without giving the physical appearance to the concerned organization. These days no business establishment or any government agency can effectively work without having a beautiful designed websites. Websites are basically the combination of multiple pages hosted on single or multiple servers for the users to access. They are broadly categorized into the static and dynamic websites. Static are those websites which are not interactive in nature whereas the dynamic websites are interactive in nature and allows the user to communicate in to and fro manner. Lots of efforts are required to produce websites which are fully user centric and can fulfill the entire needs of the users. It has become obligatory these days to launch website by each and every organization so that its reach can be global. Numbers of websites are available these days but it has been found that they do not follow the user criteria during its design phase which ultimately leads to the loss in an organization.

Different standards have been recommended for the design and development of any web-based systems and these guidelines are regularly being revised by the organization working in this field. W3C is considered to be the most effective organization in this filed which regularly publishing the standards and guidelines for designing different types of websites. Authors in this paper have tried to study the different types of W3C guidelines recommended for designing websites and have selected one of the most important parameter which effect the overall functioning of the website to a greater extent. W3C guidelines consider feedback as one of the most effective parameter which helps the users to be more interactive with the organization and also helps the organization to render service strictly as per the user requirements. Authors have given an attempt to automate the process of finding the presence and absence of feedback facility in a website. The working principle of the automated tool/parser designed for evaluating different websites is given below in Fig. 1.

The websites URL will be given as input to the interface developed and it will be supplied to the concerned server for the retrieval of HTML code for parsing and comparisons to test for the presence of FEEDBACK option in the given website.

drksujatha23@gmail.com

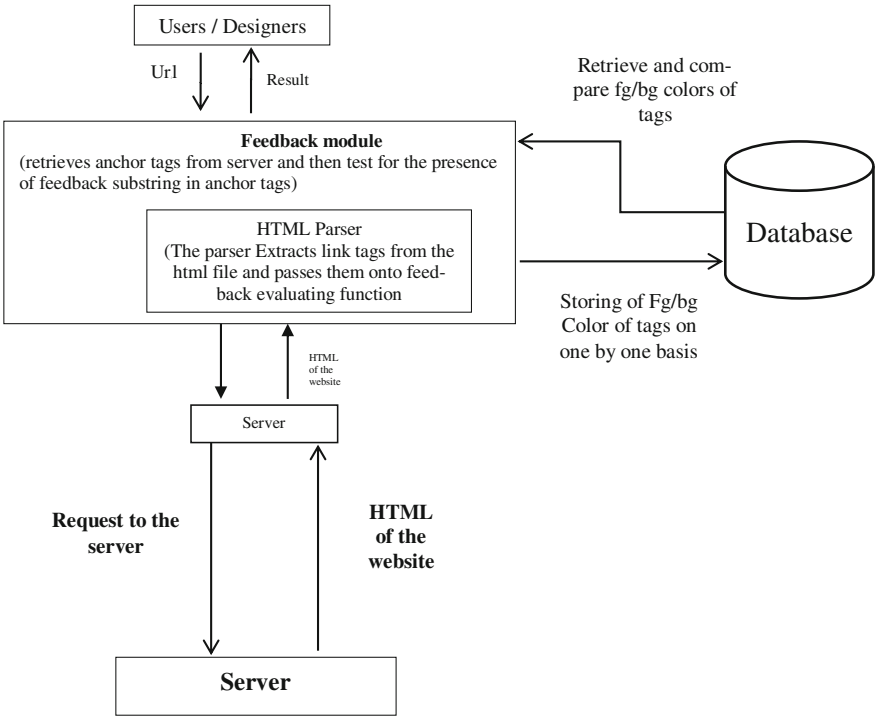


Fig. 1 Block diagram of feedback parameter

Steps adopted by an automated tool for parsing and to supply results accordingly:

Algorithm to determine the presence and absence of feedback facility in a given website.

Input to the parser: Website URL

Output given by the interface: Presence or absence of feedback facility in a given website.

Method adopted by the interface

Begin

Step I: Generates request to obtain HTML file from the web server of the concerned URL.
Step II: From HTML file all anchor tags/link tags are determined.
Step III: For every anchor tag below given steps are followed

(i) Retrieve the character string (str) within the given tags.
(ii) Check for the availability of substring "Feedback" inside the string str.

drksujatha23@gmail.com

Step IV: If for any (str contains ("feedback"))

Presence of Feedback facility is recorded in a given website
Else
Absence of Feedback facility is recorded in a given website
End

2 Methodology

2.1 Problem Identification

Website design plays a very important role in any web system design and development. Numbers of different parameters are there that influence the overall design of the websites. Out of various website design issues the parameters like sitemap [1], page loading speed [2], and feedback play significant role. While designing websites these identified parameters are taken care of so that user-centric design can be produced. In our study out of the identified parameters, we have taken feedback as one of the most important parameter that effectively helps users to communicate with organization through websites. In the current era of information technology, the feedback has become has become very important parameter to be taken into consideration while designing and executing any type of task.

Feedback assist users with the information they needed to understand about the particular organization and can also help the organization to have a better understanding about the user interacting with that organization through their websites. Feedback also helps the user to plan for the next proceeding to the next activity [3]. Providing feedback facility to the customer helps the organization to engage him for the longer period which ultimately leads to the profit-making.

Availability of feedback form for the user helps them in providing different types of comments and valuable suggestions about the website [4]. It indirectly leads to the essential part of effective learning.

Feedback is one among the eight golden rules provided by Shneiderman for guiding developers for a user-friendly interface [5]. Therefore, the need of providing feedback option in a website is of vital importance and must be included during its design and development.

2.2 Online Tool for Testing Websites

In this study, an automated tool was highly required to carry out the investigation. It will be a time consuming and hectic process to manually identify the presence and absence of feedback facility on websites. For the purpose, we designed a tool by using Dot Net framework in C#. The tool works on the latest platform and can

drksujatha23@gmail.com

Table 1 Sample data

Name of the website categories	Government	Educational	Commercial	Social websites	Job portal
Total number of websites undertaken for evaluation in each category	20	20	20	20	20

retrieve the HTML code for parsing within no time thus by giving the absence and presence of feedback facility. The tool requests the server for HTML code and once received renders the entire code for results. It has a beautifully designed interface which is user friendly in nature and can be used by layman users easily. The entire interface only demands the URL of the website for investigation. The designed tool strictly works under the recommended guidelines by W3C. The feedback feature is determined by parsing all the link tags present in a website, retrieving the strings present inside these parsed tags and then searching for the "feedback" substring for a match within each string retrieved for the parsed link tags.

2.3 Sample Data

Below given is the sample data undertaken for the study in Table 1. A total of 100 websites in five different categories of the websites were undertaken for the consideration. In each category, only 20 websites of wide variety were considered both at national and international level helping us to study for diverse types of websites. Moreover, such kind of study needs a diverse kind of websites to have a broader spectrum of the evaluations.

3 Results and Discussion

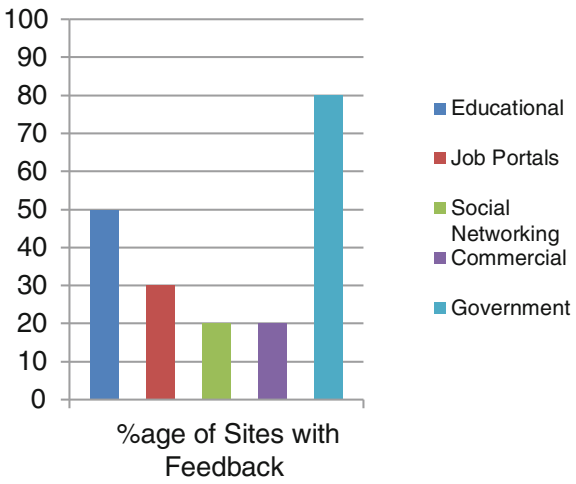
The sample data undertaken for the study in Table 1 was evaluated by the application of automated tool designed for the purpose. The websites URL was feed to the automated tool on individual basis and the results are recorded accordingly regarding the presence and absence of feedback facility on different websites. After obtaining the results, the statistical analysis is drawn in Table 2 to understand the overall scenario. Moreover, a pictorial representation in Fig. 2 was also plotted in order to have a broader spectrum and clear view of the investigative study. It has been observed that most of the websites' categories do not provide feedback facilities on their websites. The government websites undertaken for the studies show some positive results thus by making the total to 80%. The rest of websites shows poor performance thus majority of them violated the W3C guidelines

drksujatha23@gmail.com

Table 2 Results (%)

Website categories	Results	
	Websites show the presence (in %)	Websites show the presence (in %)
Educational	50	50
Job portal	30	70
Social networking	20	80
Commercial	20	80
Governments	80	20

Fig. 2 Graphical analysis



recommended for designing websites. The government websites follow the maximum of the standards recommended by W3C and prescribed in W3C guideline document WAI 1.0 [6] as far as feedback parameter is concerned.

4 Conclusion

It has been observed that majority of the websites designed in five different categories of the websites undertaken for the study fails to provide feedback facility on their websites. As consequence, the users are not able to effectively communicate with the concerned organization. It may lead to the decrease in user base which ultimately results in the organizational loss. In order to improve upon the usability factor such things are to be considered on top priority by the webmasters. Feedback plays a pivot role for effective communication between users and the organization to understand each other in more better way.

drksujatha23@gmail.com

Feedback helps in adding clarity to the website by building a one-to-one relationship between the users and the developers hence removing the gaps between them. Feedback helps in promoting the intentional behavior of the websites [7]. Its inclusion in the websites will help in understanding the purpose of being on a webpage and hence will help in removing misconceptions.

5 Limitations

The tool strictly works under the HTML environment. It fails when we determine the presence and absence of feedback facility in XML. Basically, it works for static conditions and produces no results in dynamic environment. Moreover, the tool follows only the W3C recommended guidelines and other standards given by different organizations are not taken into consideration for evaluations.

6 Future Scope

Different flaws identified in due course of time shall be taken into consideration for the next study. A number of samples shall also be increased to have a broader spectrum of the user behavior. XML will also be included in the interface. More number of parameters affecting website design shall be taken into consideration in for better website design.

Acknowledgements We are thankful to those who one or the other way supported and guided us in providing their valuable feedbacks and suggestions in the design and development of this interface. The students working in different organizations as developers guided us a lot in understating the real user scenario which was of great help to us. Academicians and scholars working in this field also contributed to a greater extent.

References

1. Manhas J (2014) Comparative study of website sitemap feature as design issue in various websites. Int J Eng Manuf (3):22–28, Published Online Dec 2014 in MECS. https://doi.org/10.5815/ijem.2014.03.03. http://www.mecs-press.net, http://www.mecs-press.net/ijem
2. Manhas J (2014) 39 Comparative study of website page size as design issue in various websites. Int J Inf Eng Electron Bus (6):33–39, Published Online Dec 2014 in MECS. https://doi.org/10.5815/ijieeb.2014.06.04. http://www.mecs-press.org/
3. Leavitt MO, Shneiderman B. Research-based web design & usability guidelines, secretary of health and human services, Professor of Computer Science, University of Maryland
4. Panta P (2009) Web design, development and security. Youngstown State University

drksujatha23@gmail.com

5. Wilkins R, Nyamapfene A. Usability driven website design—an equine sports case study, School of Engineering, Computing and Mathematics, University of Exeter, UK
6. http://www.w3.org/TR/wai-aria/
7. Wu Y-L, Tao Y-H, Yang P-C. The discussion on influence of website usability towards user acceptability, I-Shou University, Kaohsiung County, Taiwan

drksujatha23@gmail.com

A Review of Sentimental Analysis on Social Media Application



Akankasha and Bhavna Arora

Abstract Social media locales (akin Twitter, Facebook, microblogs etc.) are a global platform to share interesting ideas or news, comments, and reviews. However, feedbacks via sharing of thoughts, feelings, and comments about various products and services become key characteristics on which business in the contemporary world rely on. These are called as sentiments on social media. An attitude, believe, or acumen driven by feeling collectively called sentiment. Sentiment analysis otherwise called as opinion mining studies individuals' sentiments pointing certain elements. Web is a resourceful place for sentiment information. Difficulty arises when the phrases containing homographs are encountered. In this paper, a brief review of work done on sentiment analysis on social media applications along with various phases and levels of sentiment analysis has been discussed.

Keywords Sentiment analysis · Opinion mining · Semantic meaning
Social media application

1 Introduction

Social Media is an electronic communication format through which people communicate with each other to share information, ideas, opinion, and messages. Millions and Billions of individuals are utilizing social system locales like Instagram, Facebook, WhatsApp, Twitter, LinkedIn and Google Plus, etc., to express their emotions, opinions. They also use them to share about their day to day life activities which leads to the collection of huge and varied kinds of data. People like to share about their experiences of a particular product in terms of reviews,

Akankasha · B. Arora (✉)
Department of Computer Science & IT, Central University of Jammu, Jammu, India
e-mail: bhavna.aroramakin@gmail.com

Akankasha
e-mail: akankashaverma1000@gmail.com

© Springer Nature Singapore Pte Ltd. 2019

V. S. Rathore et al. (eds.), *Emerging Trends in Expert Applications and Security*,
Advances in Intelligent Systems and Computing 841,
https://doi.org/10.1007/978-981-13-2285-3_56

drksujatha23@gmail.com

likes, and posts and these provide a platform for the organizations to collect this data and analyze them to know about the popularity of their product and services. This process is often termed as Opinion Mining. People are sharing their views, thoughts, and opinion on different aspects every day. Internet has rehabilitated the means people utter or express their perspectives and opinions. Now it is essentially done through posts on blogs, online mediums, artifact audit websites, and so on. Social media is producing vast quantity of sentiment rich dossier as Tweets, blog posts, comments, appraisals, and so on.

Social media applications are used extensively as a source for analyzing the behaviors and opinions including sentiments of the users. Numerous social media podiums include Facebook, WhatsApp, Twitter, and Instagram etc. are common sources of getting data for the analysis process. For example, Comments, likes and stories in the case of Facebook and tweets, reviews, likes for twitter. Social media a "what's-happening-right-now" tool which lets concerned parties to take after users' thoughts and annotation on events happening in their lives in real time and immediately makes these expressions to be available in a data stream, which can be excavated utilizing proper stream mining methods [1].

This paper is organized in four sections. After the introduction to social media in Sect. 1, Sect. 2 describes the review of literature of the recent trends in the scrutiny of social media applications. Sentiment Analysis and its various phases along with its levels are discussed at length in Sect. 3. Section 4 outlines the comparative analysis of researches on sentimental analysis. To sum up, conclusion and future work are presented in Sect. 5.

2 Literature Review

Though, a lot of research going on for analysis of words in social media, but there is still a lot to contribute in analyzing of words. As proposed in [2], the author has analyzed that twitter contains sentiment rich data, that data can be analyzed using Twitter API which collects data from Twitter. It provides an interactive automatic system which predicts the sentiment of the review/tweets of the people posted in social media using Hadoop. Most of the sentiment analysis structures use bag-of-words approach for excavating sentiments from the online reviews and social media data. Instead of seeing the whole document for analysis, the bag-of-words method ponders only individual words and their count as the feature vectors. Conventionally used machine learning algorithms like Naive Bayes, Maximum Entropy, and Support Vector Machine (SVM), etc., are extensively used to resolve the problem of classification. The author in [3] has analyzed the various limitations of the existing system that can lead to biased results.

In [4], the author has proposed an automatic tool that can be used to extract words, post, and comments so that users' reviews towards the issue can be determined into happy, unhappy, or emotionless. The system implementation is performed using Java programming Language. The data is taken from Facebook as

drksujatha23@gmail.com

Facebook allow people to have their account and can comment, express feelings. Also, Facebook Query Language (FQL) is used in the JSON library to get user ID and comments. In this sentiment analysis, users' emotions are discussed.

Instagram is a social media application platform for sharing photos or it is a platform which is popular for mobile sharing application. The author in [5] has performed visualization project on Instagram data along with relationship between likes and the hashtags, location, and filters.

3 Sentiment Analysis

Sentiment analysis is a process which analyze, excerpt and can do computerizes/ automatic mining of attitudes, views, feelings, and emotions from writing, discourse/speech, chirps with the help of Natural Language Processing (NLP). It classifies opinions (text) into classes such as positive, negative or neutral [6]. Sentiment Analysis is processed to convert instructed data into meaningful information. A lot of effort has been done in "Sentiment Analysis on Twitter" by numerous researchers in recent years [7].

3.1 Phases of Sentimental Analysis

Sentimental Analysis consists of different phases as presented in Fig. 1.

Sentiment Analysis can be done in the following phases [8, 9]:

3.1.1 Data Collection: Data collected by social media applications is very huge, of varied kind and unstructured. People express their sentiment on different forums through product review, blogs, opinions, and feelings in different ways either in form of text, emotions, or sometimes even in short forms. To process and analyze data manually is a difficult task. Twitter Streaming API, twitter Oauth and programming languages like R are used to collect data from Twitter.

3.1.2 Parsing: After the data is collected, the next step is to break down data into smaller blocks or chunks. These blocks or chunks can be easily explained and managed. Sentiment Analysis Parser takes the text as input data and produces output in form of sentiments. In initial stage, it provides sentiments in binary classification (either in positive or negative). Later version allows the user to train model either on a binary dataset or in the categorical dataset (like anger, love, sad, likes, etc.)

3.1.3 Preprocessing of dataset: Millions and Billions of people post their review, opinions on Twitter and these ongoing messages called Tweets. The twitter dataset utilized is categorized into two classes, i.e., positive polarity and negative polarity and hence the sentiment exploration of the data used to

drksujatha23@gmail.com

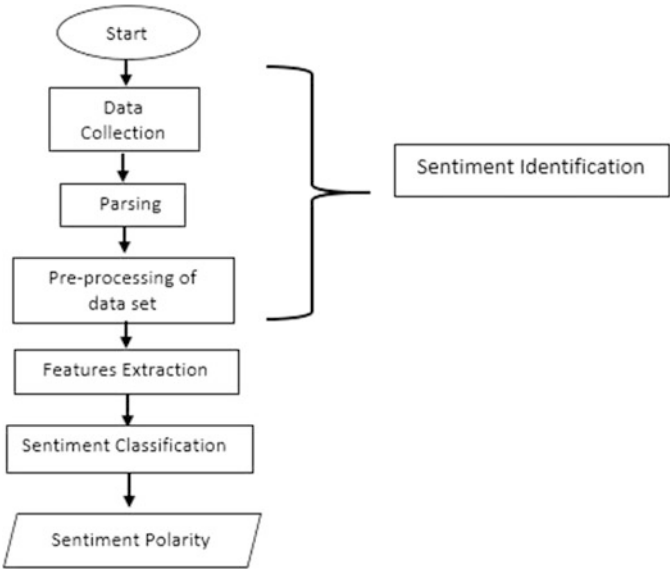


Fig. 1 Sentiment analysis architecture

detect the impact of different features. The raw data having polarity is enormously vulnerable to incongruity and redundancy. Preprocessing also called text cleaning of tweet includes subsequent facts:

Eliminate URLs targets and hashtags (For example: http://www.abc.in, @amit, #text)

Spelling correction
Swap all the emojis with their mawkishness.

Eradicate all numbers, special characters, punctuations, and symbols
Eliminate Stop Words (the, that, take, who, is, at, which, etc.)
Remove Non-English Tweets

3.1.4 Feature Extraction: In this method, relative information is extracted from the dataset which was processed in the previous phase. Further, this information is taken into consideration to calculate the positive and negative polarity in a stanza which is important to decisive the opinion of the individuals [6]. The key topographies like transcript or documents are considered as feature vectors which are utilized for the classification chore. A few illustrations features are:

Words with Frequencies: Models like Size 1 (Unigrams), Size 2 (bi-grams) and Size n (n-gram) along with their frequency computations can be used for extracting word frequency feature from the tweets.

drksujatha23@gmail.com

Parts of Speech Tags: Finding nouns, pronoun, verbs adverbs, adjectives, conjunctions, preposition, and interjections from content are virtuous indicators of opinion.
Opinion Words and Phrases: Idioms and phrases can also be utilized as features, e.g., "Action speaks louder than words".
Negation: Negation or negative words is an essential but problematic feature. The occurrence of a negation more often than not changes the polarity of the opinion, e.g., I am not happy is equivalent to sad [10].

3.1.5 Sentiment Classification: Sentiments classified into positive, negative, or neutral. Each subjective sentence is classified into good, bad like dislike, etc. Classification can be done by following methods.

Naive Bayes: It is a probabilistic classifier which predicts whether an agreed feature set fits a particular tag. Give "d" fortuitous to be the tweet and c* is a class that is allocated to d

$$C^* = \arg mac_c P_{NB}(c|d)$$
$$P_{NB}(c|d) = \frac{(P(c)) \sum_{i=1}^{m} p(f|c)^{n_{l(d)}}}{p(d)} \quad (1)$$

Naive Bayes [6]

where f is feature and m depicts the number of features. P(c) and P(f|c) are parameters computed through Maximum Likelihood Estimates.
Maximum Entropy: The classifier constantly attempts to escalate the entropy of the system by assessing the conditional dissemination of the labeled class. It also holds an overlap feature and is identical to the logistic regression technique which depicts the distribution over classes. The model is epitomized as

$$P_{ME}(c|d,\lambda) = \frac{\exp\left[\sum_i \lambda_i f_i(c,d)\right]}{\sum_c \exp\left[\sum_i \lambda_i f_i(c,d)\right]} \quad (2)$$

Maximum Entropy [6]

where c defines class, d depicts tweet, and λi is the weight vector. The role of weight vectors is to decide the importance of a feature in cataloguing.
Support Vector Machine (*SVM*): SVM examines the data, defines the decision margins, and uses the kernels for computation which implemented in input space. SVM additionally coordinates classification and regression which are valuable for statistical erudition theory and it likewise aids perceiving the factors accurately, that should be considered, to comprehend it successfully [6].

drksujatha23@gmail.com

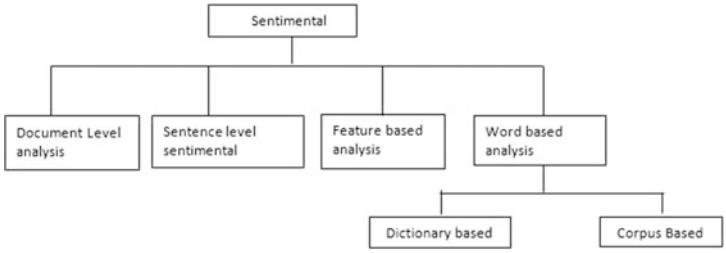


Fig. 2 Levels of sentiment analysis

3.2 Levels of Sentimental Analysis

Sentiment analysis is divided into various levels according to the terms under consideration. The pictorial representation is given in Fig. 2 [6].

a. Document level: It is the simplest form of classification. In this, entire document is classified one or the other class that is positive or negative class [11].
b. Sentence or phrase level: It classifies sentence into 'positive', 'negative', or 'neutral' polarity. It identifies subjective or objective sentences.
c. Aspect level or Feature level: Recognize and excerpt object features that have been mentioned on by an opinion holder. For example 'My DSLR photograph quality is worthy but memory is low'. Here, DSLR and Quality display positive sentences but memory display negative sentence [12].
d. Word Level: Latest work have utilized the earlier schism of words and phrases for sentiment arrangement at sentence and manuscript levels. Word level sentiment classification uses most parts of adjectives as features same as adverbs. Two approaches of commenting on sentiment at the word level are:

   Dictionary-Based Approaches: This method is built on the acceptance of items that are normally collected and illustrated manually or physically. It is done by searching the synonyms and antonyms of a vocabulary.
   Corpus-Based Approaches: It is the study of real-life languages in the form of text (structured text) or speech (audio files).

4 Comparative Analysis

This section presents a brief summary and comparison of work that has been by researchers in this field. It can be seen that the work based on sentiment and opinion mining is recent and uses the latest techniques and software for its analysis. Table 1 presents a brief comparison on various social media applications along with the tools and algorithms that have been used to analyze sentiments and opinions.

drksujatha23@gmail.com

Table 1 Comparative analysis of researches on sentiment analysis

	Trupthi and Pabboju [2]	Jose and Chooralil [3]	Zamani et al. [4]	Chang [6]
Year	2017	2015	2013	2016
Social media application	Twitter	Twitter	Facebook	Instagram
Tools/ Algorithms used	Hadoop, MapReduce	Naive Bayes, Maximum Entropy, SVM	Java, Facebook Query Language (FQL) is used in the JSON	Python
Analysis	Real-time sentimental analysis on the tweets that are extracted from the twitter and provide time-based analytics to the user	Sentiment classification of Twitter messages using lexical resources SentiWordNet and WordNet along with Word Sense Disambiguation	Opinion mining and sentiment analysis components for extracting both English and Malay words on Facebook	Analysis of data from Instagram. It also analyzes hashtags in different locations to showing visual culture difference between different cities

5 Conclusion and Future Scope

Social Media is full of smart people who share their experiences, reviews and thoughts, all for free. In the contemporary world, the life of internet user is discontented without constant link to social media. One of the major factors that influence the social media is the way the users express their sentiments online. The sentiments that the users express online will serve as a major parameter in using the social media in the context of user behavior and polarity (positive, negative, or neutral). The sentimental analysis is a tool that can be used to express one's feelings online. The analysis of these becomes a challenging task where homograph is used. This paper discusses sentiment analysis, its phases, and levels along with techniques. It will only work for English language but in future it can be used for Multilingual. Also, various opinion summarization algorithms can be applied to generate a summary of reviews given by users.

Note: This is a baseline for an ongoing M.tech Research project on sentimental analysis on social media homographs.

drksujatha23@gmail.com

References

1. Jin Z, Yang Y, Bao X, Huang B (2016) Combining user-based and global lexicon features for sentiment analysis in Twitter. IEEE. 978-1-5090-0620-5
2. Trupthi M, Pabboju S (2017) Sentiment analysis on Twitter using streaming API. IEEE. 978-1-5090-1560-3
3. Jose R, Chooralil VS (2015) Prediction of election result by enhanced sentiment analysis on Twitter data using word sense disambiguation. IEEE. 978-1-4673-7349-4
4. Zamani NAM, Abidin SZZ, Omar N, Abiden MZZ (2013) Sentiment analysis: determining people's emotions in Facebook. ISBN 978-960-474-368-1
5. Chang S (2016) Instagram post data analysis. arXiv:1610.02445v1
6. Kharde VA, Sonawane SS (2016) Sentiment analysis of Twitter data: a survey of techniques. Int J Comput Appl 139(11)
7. Arora D, Li KF, Neville SW (2015) Consumers' sentiment analysis of popular phone brands and operating system preference using Twitter data: a feasibility study. IEEE. 1550-445X
8. Godsay M (2015) The process of sentiment analysis: a study. Int J Comput Appl 126(7)
9. Medhat W, Hassan A, Korashy H (2014) Sentiment analysis algorithm and applications: a Survey. Ain Shams Eng J 5
10. Eshleman RM, Yang H (2014) A spatio-temporal sentiment analysis of Twitter data and 311 civil complaints. IEEE, 978-1-4799-6719-3
11. Gokulakrishnan B, Priyanthan P, Ragavan T, Prasath N, Perera A (2012) Opinion mining and sentiment analysis on a Twitter data stream. IEEE, ICTer, pp 182–188
12. Grandin P, Adán JM (2016) Piegas: a system for sentiment analysis of tweets in Portuguese. IEEE Lat Am Trans 14(7)

drksujatha23@gmail.com

Trust Prediction Using Ant Colony Optimization and Particle Swarm Optimization in Social Networks



Rajeev Goyal, Arvind K. Updhyay and Sanjiv Sharma

Abstract Today is the world of virtual communication, development of online social network facilitates the users to communicate and collaborate with each other via several tools available on online social networks. Users share their experience, relation, views, etc., on the online social network. It is very important to provide trust mechanism to establish a trust relationship between users and the source of information and the consumers of the information on the social network. Trust prediction has become one of the most important tools for finding and identifying the potential trust relationship between any online communities. Such a reliable source of information or the users in the community would be recommended to other targets and online communities. This paper has proposed a new method to provide trust prediction through a hybrid approach of the ant colony optimization algorithm and particle swarm optimization algorithm. The proposed method can give a more efficient result by improving the process of pheromone update by particle swarm optimization.

Keywords Trust · Link prediction algorithm · PSO · ACO · OSN

1 Introduction

Online social network is like a virtual world that provides the freedom of having in touch or connected to the communities. Boyd et al. [1] outlined that the online social network is the web-based application that provides a person to build a private

R. Goyal (✉) · A. K. Updhyay
Amity University Madhya Pradesh, Gwalior, Madhya Pradesh, India
e-mail: rgoyal@gwa.amity.edy; goyal.rajeev@gmail.com

A. K. Updhyay
e-mail: akupdhyay@gwa.amity.edu

S. Sharma
Madhav Institute of Technologies and Science, Gwalior, India
e-mail: er.sanjiv@gmail.com

© Springer Nature Singapore Pte Ltd. 2019

V. S. Rathore et al. (eds.), *Emerging Trends in Expert Applications and Security*,
Advances in Intelligent Systems and Computing 841,
https://doi.org/10.1007/978-981-13-2285-3_57

drksujatha23@gmail.com

and public profile within a restricted system, provide the list of trusted users to whom the user can share information, and traverse and sight the uses connection in the online social community made by others. Growth and acceptance of the virtual communities of the online social network are growing day by day. Widespread and extensive uses of online social application raise the pursuit and outline of activities of each user. Zhang et al. [2] proposed that the quality of system improves and examine online social networks framework.

An immense number of users join communities and share their thoughts, information and develop their relationships in online social networking sites such as Twitter and Facebook. As a result of this, providing trust between the user's connection from source information to target information and among all the users in a network is one of the big issues. Trust prediction between users is calculated by multiple contexts like that role impact factor, reliability, preference similarity, social intimacy, and existing trust. Artificial neural network (ANN) is one of the heuristic methods for the prediction of trust between users in the online social network. Ant colony optimization (ACO) is also a heuristic algorithm that provides a solution of several standard problems of computer science, such as traveling salesman problem, machine learning, network routing, graph color, job sequencing, etc.

1.1 Swarm Intelligence

Birds Fish, Ant, Bee are the small species but they perform their tasks such as finding food by their collaborative behavior, communication, and decentralization. They inhabit a complex group and perform the very complex task in a much optimized way.

Several types of study have been done in fields of science like in computer science, where researchers apply different swarm intelligence techniques such as ant colony optimization, particle swarm optimization to solve such complicated problems.

1.2 Ant Colony Optimization (ACO)

Small species like ant are mostly blind but with the help of communication and group support, they find the food. They use a special chemical called pheromone to communicate within the group and communicate about the source of food. With the help of pheromone, they take the decision where a nest has to move to the food or from food to the nest. While locating food, ant spreads pheromones on to the path and with this, the other ant creates a trial. This behavior of the ant is investigated by several researchers. One of the methods named double bridge which shows the mathematical representation of decision-making through ACO; in this method nest and food are denoted by nodes and path is denoted by the arc or edge.

drksujatha23@gmail.com

Corresponding to each path probabilistic rules are defined for movement of vertices and help in decision making.

1.3 Particle Swarm Optimization (PSO)

Eberhart and Kennedy [3] have developed a technique for optimization known as particle swarm optimization (PSO) motivated by the birds who search for food in the sky randomly. In this technique, birds search for the shortest path to the food by following the other birds. PSO is similar to the genetic algorithm (GA). PSO is based on population and searches optimum path by updating generations. In PSO particle is the optimum solution for searching an optimum solution, i.e., the smallest path.

PSO is easy to implement and have less number of the parameter to implement. PSO can be used in different fields such as social network, fuzzy logic, neural network, etc. Steps of PSO is as follows.

In the first step, initialize a group of particles and then find the optimum solution by updating the solutions. In the loop, each one of the particles is initialized with two values; the first one is called best solution, i.e., fitness value also known as fbest. And, the second one is searched by the PSO and this value is the best value obtained by any of the particles. This value is known as gbest or global best.

After searching these values, the particle changes their position and velocity as follows.

$$\text{val}[] = \text{val}[] + \text{x1} * \text{rand}() * (\text{fbest}[] - \text{pre}[]) + \text{x2} * \text{rand}() * (\text{gbest}[] - \text{pre}[]) \quad (1)$$

$$\text{per}[] = \text{per}[] + \text{val}[] \quad (2)$$

here particle velocity is val[], current particle per[].Random number between (0,1) rand ().learning factors are x1 and x2. Mostly x1 = x2 = 2.

Particle swarm optimization is used in several applications to find the optimized solution. In this research, it is used to improve the pheromones amount.

2 Related Works

In online social network, trust is one of the most important contexts to predict links and taking the decision to obtained any service or relation. Providing a trustworthy environment for consumers and service providers in the online social network is one of the most challenging tasks. Golbeck and Hendler [4] give trust model that search average trust value on to the trust of the social path. Guha et al. [5] have proposed a model for trust propagation, which computes the value of trust between the source (service provider) and destination (consumer) by the number of hopes in

drksujatha23@gmail.com

trust propagation. Walter et al. [6] established the online social networking recommendation system which recommends behavior of the link (member of the online social network) by allocating a trust value on the base of priority. Jamali and Ester [7] established a method for service provider and consumer to perform a random walk between the nodes and search rating given by the consumer. The service provider wants to trade his service or product which is preferred by the consumer. Researchers calculate the degree of confidence for the service provider with the help of service providers rating, nodes, and a number of random walks. Bedi et al. [8] investigated that the trusted user always carried out with the recommendation. Kuipers et al. [9] proposed a heuristic algorithm for the selection of optimal path with multiple contexts also known as Social Trust Path Selection Algorithm (H-MOCP). Abbasimehr and Tarokh [10] proposed a new way to predict the trust in online communities using neuro fuzzy and predict the trust among the online social network through neural network.

Yu et el. [11] give a method for computation of service for k path for intermediate node and had compacted the space for solution search and also decrease execution time known as MCSP−K.

Van Wang et al. [12] proposed H-OSTP based on the Dijkstra algorithm for finding the optimal social trust path which had also used the trust aggregation and impact factor of trust and used social pheromones. One of the enhanced version of H-OSTP, i.e., MFPB-HOSTP-MFPB HOSTP used three-part first search backward path in local network from source to destination. Second, find a forward path in the local network for the intermediate path. And then composite path which has the quality of trust attribute and highest aggregation.

3 Proposed Algorithm

For the searching trust between nodes, a heuristic method such that ACO is used [13]. It has a very high complexity and this is one of the biggest drawbacks because it involves volume calculation. To overcome this problem, a new hybrid algorithm has been proposed ant colony optimization and particle swarm optimization. Here the designated graph (weighted) contains n + 1 iteration.

The process to search for trust in an online social network using hybrid ACO and PSO is that in every repetition, the best solution to the problem of the available solution is found. Taking into consideration the mentioned items, the general routine of the proposed algorithm can be summarized as follows:

Step 1. Parameter Initialization
Step 2. Repeat steps 3–5 until finding the desired solution.
Step 3. Graph traversal by ants and subgraphs creation as trust value solution
Step 4. Calculate the obtained trust value
Step 5. Pheromones Update by PSO
Step 6. End.

drksujatha23@gmail.com

Path selection of ant is done by the pheromone spread in the available path [14]. Higher pheromones mean there is a high probability to select the path. In ACO at time t(0, 1, 2…), path x(i, j) pheromone is represented by pij(t).at start t = 0 the pheromone is initialized with +ve value let us take A means Aij(0) = A. At time t = 0, ant let us say Q starts and stop on 1. After that every ant a path and shift to next stop on the basis of pheromone. This process lasts till ant search the food. The probability to shift an ant from one stop (a = 1, 2, 3,…x) to (a = 1, 2, 3…) at t time t is as follows:

$$P_{ij}^{x}(t) = \begin{cases} \frac{\left[A_{ij}(t)^{\alpha}\left[u_{ij}\right]^{\beta}\right]}{\sum_{f\in J_k(i)}\left[A_i f(t)^{\alpha}\cdot\left[u_{ij}\right]^{\beta}\right]} & , If\ j \in S_k(i) \\ 0,\ otherwise \end{cases}$$

Here, u ij is the utility percentage in OSN. Bigger the nij ant select j node. Set of path let say Sx(i) at node x. so the path for the nodes correspond to selected ant X is Xx = {Xi| i = 1, 2, 3…}.

After the upgradation of pheromones, the equation is as follows:

$$\Delta A_{ij}^{x} = \begin{cases} \frac{R_x}{Q} \text{ if ant x passes path p[i,j]} \\ 0,\ otherwise \end{cases}$$

Here, ΔA_{ij}^{x} is the pheromones path on p (i, j) Rk is the solution for the current iteration.

$$P_i' = \begin{cases} \frac{1-u_i}{\sum_{j\in L'_{best}}(1-u_i)} & , If\ j \in i'_{best} \\ 0,\ otherwise \end{cases}$$

4 Summary and Result

The result of an experiment on epionoin.com dataset (Fig. 1). Hybrid ACO and PSO is an iterative algorithm a Bi Net and SCAN has. Their result has improved with each iteration.

Methods	Min	Mean	Max
Hybrid ACO and PSO	0.65	0.71	0.75
Bi Net	0.58	0.64	0.70
SCAN	0.54	0.57	0.60
FDRS	0.49	0.53	0.58

drksujatha23@gmail.com

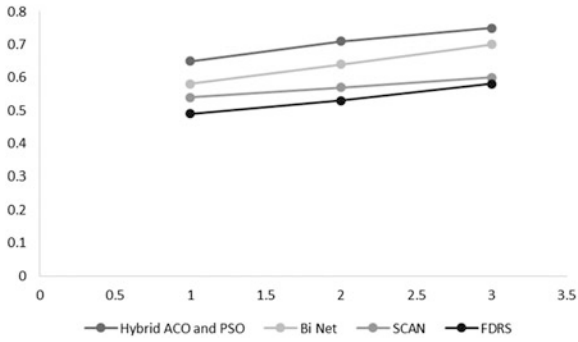


Fig. 1 Result on the dataset (epinion.com)

The result has been summarized as follows.

- Ant colony optimization is one of the solutions that is able to search and deliver in whole solution space.
- Path selection allows the ant to try multiple times and multiple paths.
- Pheromone update method uses PSO to increase efficiency and remove complexity.

5 Conclusion

The paper has proposed a novel approach for calculation of trust in the online social network. The approach has used a hybrid of ACO and PSO for searching the best solution. Ant colony optimization is used for trust calculation and particle swarm optimization is used for searching swarm pheromones. Both of this algorithm is used to find trust or a path among the nodes, by using hybrid approach trust calculation is more efficient. This novel approach provides a new way to solve trust-related issues.

References

1. Heer J, Boyd D (2005) Vizster: visualizing online social networks. In: IEEE Symposium on information visualization, 2005. INFOVIS 2005. IEEE
2. Lin H, Fan W, Zhang Z (2010) A qualitative study of web-based knowledge communities: examining success factors. E-collaboration technologies and organizational performance: current and future trends 214
3. Eberhart R, Kennedy J (1995) A new optimizer using particle swarm theory. In: Proceedings of the sixth international symposium on Micro machine and human science, 1995. MHS'95., IEEE
4. Golbeck J, Parsia B, Hendler J (2003) Trust networks on the semantic web. In: Cooperative information agents VII, pp 238–249

drksujatha23@gmail.com

5. Guha R et al (2004) Propagation of trust and distrust. In: Proceedings of the 13th international conference on World Wide Web. ACM
6. Walter FE, Battiston S, Schweitzer F (2008) A model of a trust-based recommendation system on a social network. Auton Agent Multi-Agent Syst 16(1):57–74
7. Jamali M, Ester M (2010) A matrix factorization technique with trust propagation for recommendation in social networks. In: Proceedings of the fourth ACM conference on recommender systems. ACM
8. Bedi P, Kaur H, Marwaha S (2007) Trust-based recommender system for semantic web. IJCAI 7
9. Kuipers F et al (2004) Performance evaluation of constraint-based path selection algorithms. In: IEEE network 18.5, pp 16–23
10. Abbasimehr H, Tarokh MJ (2015) Trust prediction in online communities employing neuro-fuzzy approach. Appl Artif Intell 29(7):733–751
11. Yu T, Zhang Y, Lin K-J (2007) Efficient algorithms for web services selection with end-to-end QoS constraints. ACM Trans Web (TWEB) 1(1):6
12. Liu G, Wang Y, Orgun MA (2010) Optimal social trust path selection in complex social networks. AAAI 10
13. Sanadhya S, Singh S (2015) Trust calculation with ant colony optimization in online social networks. Procedia Comput Sci 54:186–195
14. Soleimani-Pouri, M, Rezvanian A, Meybodi MR (2012) Finding a maximum clique using ant colony optimization and particle swarm optimization in social networks. In: Proceedings of the 2012 international conference on advances in social networks analysis and mining (ASONAM 2012). IEEE Computer Society
15. Gambhir S, Kumar V (2015) Bidirectional trust calculation in online social networks. In: 2015 4th international conference on reliability, infocom technologies and optimization (ICRITO) (Trends and future directions). IEEE
16. Yadav A, Chakraverty S, Sibal R (2015) A survey of implicit trust on social networks. In: 2015 international conference on green computing and internet of things (ICGCIoT). IEEE

drksujatha23@gmail.com

Stock Market Price Prediction Using LSTM RNN



Kriti Pawar, Raj Srujan Jalem and Vivek Tiwari

Abstract Financial Analysis has become a challenging aspect in today's world of valuable and better investment. This paper introduces the implementation of Recurrent Neural Network (RNN) along with Long Short-Term Memory Cells (LSTM) for Stock Market Prediction used for Portfolio Management considering the Time Series Historical Stock Data of Stocks in the Portfolio. The comparison of the model with the traditional Machine Learning Algorithms—Regression, Support Vector Machine, Random Forest, Feed Forward Neural Network and Backpropagation have been performed. Various metrics and architectures of LSTM RNN model have been considered and are tested and analysed. There is discussion on how the sentiments of the customer would affect the stocks along with the changes in trends.

Keywords Recurrent neural network · Long short-term memory
Trading · Portfolio optimization

1 Introduction

Predictions on the stock market have been considered as an important study object for many decades [1]. But its complexity and dynamic environment have been proven it to be a very difficult task [2, 3]. Predicting price and trend of the stock market are the indispensable aspects of investment and finance. Many researchers have worked and proposed their ideas to forecast the market price to make a profit while trading using various techniques such as technical and statistical analysis.

K. Pawar · R. S. Jalem · V. Tiwari (✉)
DSPM IIIT, Naya Raipur, India
e-mail: vivek@iiitnr.edu.in

K. Pawar
e-mail: kriti15100@iiitnr.edu.in

R. S. Jalem
e-mail: raj15100@iiitnr.edu.in

© Springer Nature Singapore Pte Ltd. 2019

V. S. Rathore et al. (eds.), *Emerging Trends in Expert Applications and Security*,
Advances in Intelligent Systems and Computing 841,
https://doi.org/10.1007/978-981-13-2285-3_58

drksujatha23@gmail.com

Observing and predicting trends in the price of the stock market is challenging because of noise and uncertainties involved. There are numbers of factors that may affect the market value in a day such as country's economic change, product value, investors' sentiments, weather, political affairs, etc. [4]. The authors have also studied and researched on the trends and behaviour of stock prices and what all factors affect prices the most. Al-Nasseri et al. [5] have analysed the divergence of opinion and the impact of the disagreement on Stock Returns and Trading Volumes.

In this research, RNN along with LSTM is used for predicting the movement of the stock market. The stock market is also mainly affected due to the sentiment of customers or buyers that is their opinion on a particular product or service provided by the company is also one of the main additions to the fluctuations in stock prices. The research also compares the RNN-LSTM model with many Traditional Machine Learning algorithms. Several possibilities have been considered for the model and the model has been tested accordingly to several possibilities and is analysed for different configurations of the model.

2 Literature Survey

Aditya Gupta and Bhuwan Dhingra in [6] used Hidden Markov Model to predict the close price of the stocks of next day. They have used historical stock prices of different companies such as the Apple Inc., IBM Corporation, TATA Steel and Dell Inc. Inputs were High Price, Low Price, Open Price and Close Price. Model for each stock was supposed to be independent of every other stock. The model was first trained for a period of 7 months. The model was tested using MAPE values.

Lin et al. in [7] have been used SVM based approach to predict the price of stock market trends. They have solved the problem in two parts, i.e. feature selection and prediction of the direction of trends in the market. SVM correlation has been used to select the features which affect the price mostly. Linear SVM is applied to the data series to predict the direction. They have shown the system to select the good feature and control overfitting on stock market tendency prediction.

Dinesh and Girish in [8] developed a model based on linear regression, as in linear regression, there is given set of input for output and by developing a model based on mathematical foundation output is predicted. They have used Open Price, High Price, Low Price and volume as input to the model and independent variable and Close Price as the label, and had considered Date is used as a variable index. By comparing the linear regression model with polynomial and RBF regression approach, linear regression has proven to be the best among both.

Yang et al. in [9] first select the most relevant features for prediction of the stock price by calculating maximal information coefficient. They build their assembler model using three different outstanding classifiers on stock market trend prediction SVM, Random forest and AdaBoost and collectively named as SRAVoting. They validate their model on Chinese Stock Market and come to the conclusion that

drksujatha23@gmail.com

SRAVoting gives higher accuracy than SVM but at the same time lesser buy/sell strategies than SVM.

In [10], they have compared Random Forest, SVM and Gradient Boosted Trees for forecasting Moroccan stock market for the short term. The empirical results showed that all the three models have given very satisfactory results and they have short time responses and hence shows that these methods can be usable for a short time. They have come up with the result that Random Forest and Gradient Boosted Trees is superior to Support Vector Machine. They have also suggested that proper feature selection and reduction is required for more accurate results.

In [11], they have used optimized ANN to predict the direction of the price movement of the next day of Japanese stock market. To improve the accuracy of the predicted direction they have introduced the Genetic Algorithm. Another method hybrid GA-ANN is also used in order to predict the direction of price movement. After comparing both methods, the second method is prone to give satisfactory result in term of accuracy. By adjusting weights and biases of ANN using GA, the model gave the Hit ratio of 86.39%. The proper feature selection is required to gain more accuracy.

The backpropagation neural network remains the universal and most fruitful prototype for multilayer networks [12]. The typical backpropagation neural network contains three layers: input, output, and should have at least one hidden layer. The networking potential for the selected size of the dataset depends on a number of neurons at each layer and the number of hidden layers to gain correct result [13].

The proposed scheme/method is to preprocess the dataset, which is followed by an altered backpropagation neural network scheme/algorithm with the attention to contemporary fashion and event. At last, the predicted value from the model is compared to the input value to minimize error. Here, we can concentrate on accuracy or, in other terms, minimize the error to get the accurate predicted value as compared to actual value. The problem with Feed-Forward neural network [14] has been overcome by supervised learning methods where prior knowledge is not required. This gives the better results than the Random Forest, we can make the layers recurrent for considering the sequential stock data.

3 Prediction Using LSTM RNN

3.1 Recurrent Neural Networks

Recurrent Neural Networks are the class of Neural Networks [15] where the units are recurrently connected. This allows them to use their internal memory for processing the sequence of inputs. This allows them to be used for handwritten recognition, text generation, the stock market or speech recognition. Recurrent Neural Networks are used in this project since long-term dependencies [16] in the data needs to be considered for the stock data. While due to the inability to store the

drksujatha23@gmail.com

memory for much amount of time, Vanishing Gradient descent problem may occur, i.e. after every iteration in the neural net, the data it holds gets vanished going deeper. Due to which Long Short-Term Memory cells instead of traditional Neuron-like cells are used.

3.2 Long Short-Term Memory Cells

Long short-term memory (LSTM) block [17] or network is an advancement to the simple recurrent neural network which can be used as a building component or block for an eventually better serial analysis using the recurrent neural network. LSTM block itself a recurrent network as it contains recurrent connections like connections as in a conventional recurrent neural network. LSTMs is designed specifically as a recurrent neural network architecture to consider them for long-term dependencies more accurately than the conventional Recurrent Neural Networks. According to [16], LSTM along with RNN have outperformed the Deep Neural Networks (DNNs) and the simple RNN models for predicting the movements in stock data or speech recognition.

Conventional DNNs can only provide modelling for a fixed sized sliding window where the network does not interdepend on the previous time steps which would so do not provide a good modelling for the stock data (Fig. 1).

Data is retrieved from the online open source financial data provider Yahoo Finance. For the training purpose, historical stock data of S&P 500 was considered as it has a large database. The data is then normalized and split into the training and the testing data. The training data is then used to train the built LSTM RNN model to train for predicting or forecasting the sequence of the stock data. Then, the model is tested on several stocks like the Apple Inc., Tesla Inc., Google and the forecasted versus the actual data is visualized through the plot in the results section (Fig. 2).

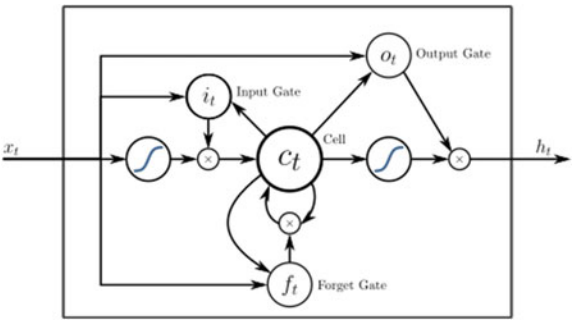


Fig. 1 Long short-term memory cell

drksujatha23@gmail.com

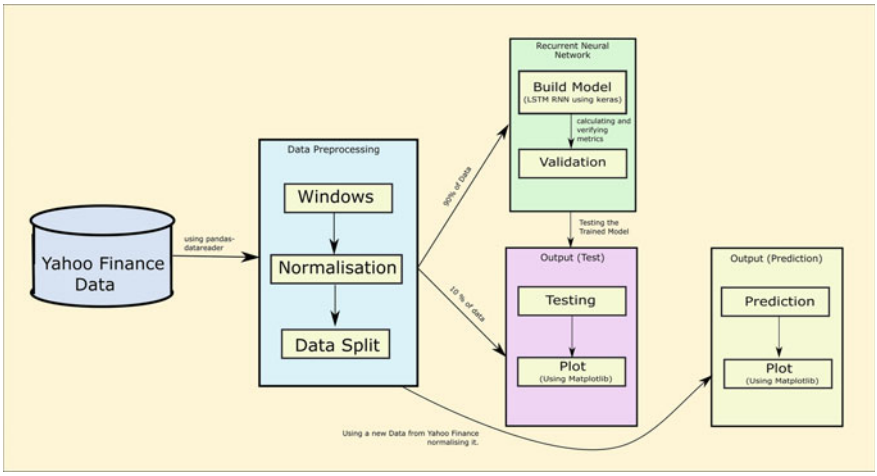


Fig. 2 Architecture of the proposed model

Algorithm: Prediction using LSTM RNN

Input: Historical Stock Prices of stocks.
Output: Predicted Stock Prices for n Data Points.

Data <= Historical Stock Data to be retrieved from **Yahoo Finance.**

Adj Close <= Adjacent Close values retieved from Data.

Function Preprocessing(Adj Close, sequence_length)

Input: Adj Close is the Adjacent Close values that are retrieved from the Data.
Output: The Data is normalised and split to Train and Test

Data_windows <= windows(Adj Close, sequence_length)

Function Normalise(Data_windows)
Normalised_data = []
For i in Data_windows:
Normalised_window = [((float(p) / float(i[0])) - 1) **for** p in i]
Normalised_data.append(Normalised_window)
Return Normalised_data
row <= 90% of the shape of normalised data
Train_data <= [: row, : -1]
Train_label <= [: row, -1] *//as sequence prediction is done the same data is divided to*
Test_data <= [row:, : -1] *train_data, train_label, test_data and test_label.*
Test_label <= [row:, -1]

drksujatha23@gmail.com

Function model(a, b, epoch, batch_size);

Input: a and b are the Train_Data and Train_label respectively.
Output: the Trained model is obtained.

Network = **sequential()**
Network.**add**(LSTM(input, output, dropout)) *//Input Layer of LSTM RNN*
Network.**add**(LSTM(cells, activation, dropout))$_1$… Network.**add**(LSTM(cells, activation, dropout))$_k$…Network.**add**(LSTM(cells, activation, dropout))$_n$ *//Hidden Layers*

Network.**add**(LSTM(output, output_activation) *//Output Layer*

Network**.compile**(loss_function, optimizer) *//Defining Optimisation of the model*

Network**.fit**(a, b, epoch, batch_size, validation) *//Training the model*

Function Plot(model, validation_data, validation_label)
Input: model is the trained model i.e., Network is given as the input, validation_data is the Test_data and validation_label is the Test_label.
Output: Plot is obtained i.e., Predicted vs the validation_label

Prediction = model.**predict**(validation_data) *//Predicting the validation_label for validation_data*
Plot(predicted vs validation_label) *//Plotting graph of Predicted vs the True Data*

4 Results

All the results are obtained by different configurations of the model with the loss function of mean squared error and an optimizer of Adam (Adam is an optimization algorithm which is used to update the weights of the network iterating based on training data). Adam was created by Jimmy Ba from the University of Toronto and Diederik Kingma from OpenAI in their paper titled 'Adam: A Method for Stochastic Optimization'.

There are several LSTM architectures we can build as shown in Fig. 3, using the LSTM structure or the combination of LSTM and RNN. In this paper, several architectures are tested to find the best model with the lowest loss value.

The loss observed for the LSTMP architecture after three epochs are 0.5770.

The loss observed for the Deep LSTM after three epochs are 3.1464e−04, which is much better than the other models.

drksujatha23@gmail.com

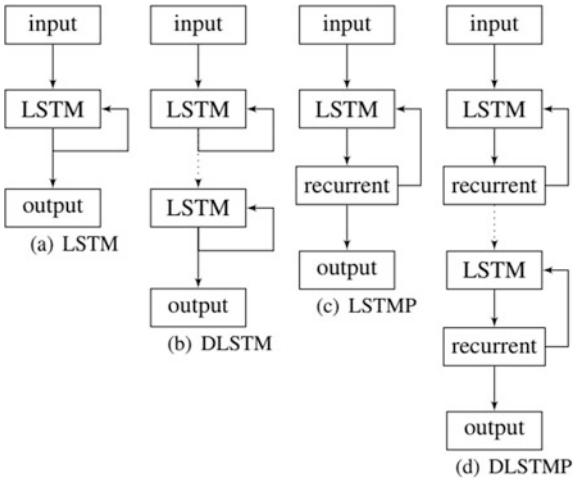


Fig. 3 Several architectures of LSTM

Table 1 Several LSTM architectures with their loss observed

	I	II	III	IV	V	VI
Cells	128	128	256	256	512	512
Layers	1	2	1	2	1	2
Loss[a]	2.554e −04	2.835e −04	2.307e −04	2.455e −04	2.080e −04	2.0540e-04

[a]The loss function used is "Mean Squared Error"

As the LSTM architecture has got the best loss value, the behaviour is also checked by varying the inner architecture, i.e. the number of cells and the number of layers keeping constant the activation function of the hidden layer as hyperbolic tangent function and of the output layer as the rectified Linear Unit function as these are observed to give the best results. All the results are obtained for five epochs (Table 1).

The historical stock data of AAPL (Apple Inc.), GOOG (Google), and TSLA (Tesla, Inc.) have been considered from the Yahoo Finance and have been normalized to fit the data into the model for prediction and the predictions have been taken for an epoch = 10 (10 iterations through the entire data and the model is optimized) with a batch size (number of data points to be considered at a time to be 128) (Figs. 4, 5 and 6).

For Epoch = 10.

drksujatha23@gmail.com

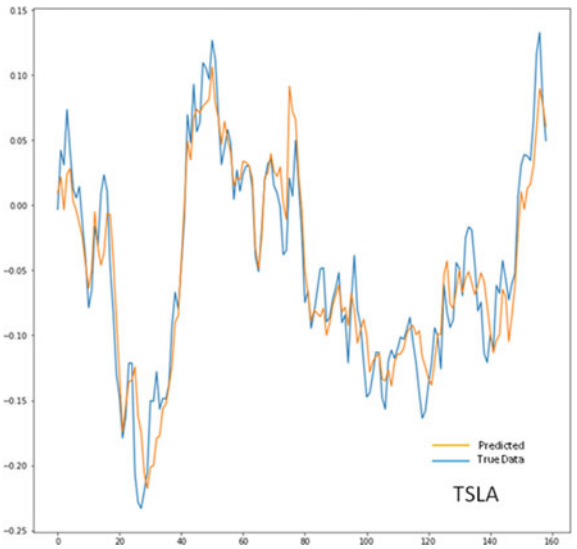


Fig. 4 Predicted versus true data movements of Tesla Inc.

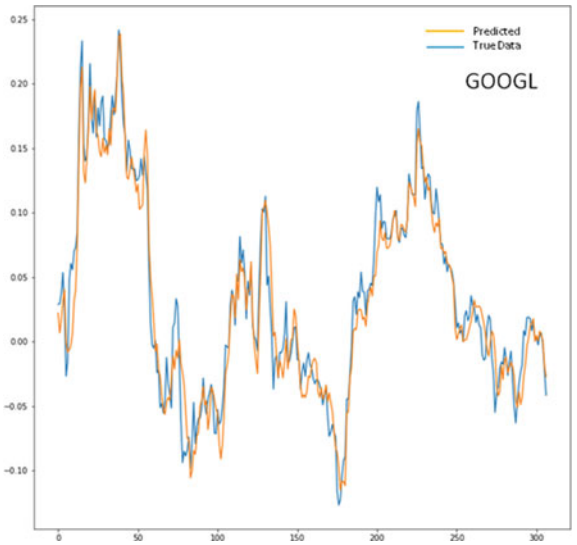


Fig. 5 Predicted versus true data movements of Google Inc.

drksujatha23@gmail.com

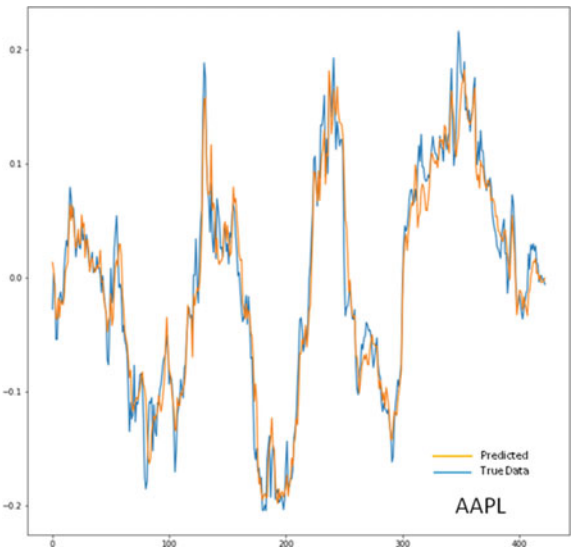


Fig. 6 Predicted versus true data movements of Apple Inc.

5 Discussion

When an investor has been predicted and understood the behaviour of the market based on rising and fall in the price of the assets. The Next problem comes in the trading strategy is how much proportion of his share he could distribute in different stocks in his portfolio [18], he should have prior knowledge before investing. The adjustment of the proportion of wealth in different stocks to gain profit is a very important factor to understand. Markowitz's portfolio optimization [19] is purely based on the mathematical foundation and gives the very satisfactory result to distribute ones share in the different market [20].

We have calculated return as

$$r = (adjClose(i) - adjClose(i-1))/adjClose(i)$$

annual return by multiplying 12 to return, the variance of return, the covariance matrix of return, weights and sum should be 1,

$$\begin{aligned} expected\ return &= transpose\ of\ weight * return, \\ volatility &= (transpose\ of\ weights * covariance\ of\ return * weights)^{1/2}, \\ Sharpe\ ratio &= expected\ return/volatility \end{aligned}$$

We then adjust the weight to gain profit and select the portfolio which has max Sharpe ratio and in volatility. By performing Markowitz Portfolio Optimization technique, we can understand how much we must invest our money in a market.

drksujatha23@gmail.com

6 Conclusion

To develop the prediction model, the implementation process should be gone through relevant data collection, data preprocessing to remove noise and missing values. Analysing the best Algorithm followed by model evaluation. The research introduced in this paper uses the Recurrent Neural Network with LSTM cells to predict the movement of stock market exchange.

The results show that RNN-LSTM model prone to give more accurate result than the traditional machine learning algorithms.

This model can be proved to be productive for individual traders as well as for corporate investors. They can get the future behaviour of market price movement and take the proper action to make a profit.

In future work, the model should be considered different features and aspects of the market to make prediction more accurate. Also, we intend to use reviews of the users on the product to predict the change in the market.

7 Future Implementation

Stock Data not only depends on the Trend in the Historical Data, it also mainly depends on the product value or the satisfaction of the customers with the company's market [21]. So, the future implementation includes analysing the sentiment of the customers reviewed on the products related to a company or its domain and add this analysis to the prediction using RNN-LSTM. Research work has been carried out using Naive Bayes Classifier with Large Movie Review Data as the Training set and then tested (sentiment analysis) on Amazon Review Data.

Thus, the results obtained through both the analysis are considered to obtain better forecasting of the portfolio.

Acknowledgements This research was partially supported by DSPM International Institute of Information Technology Naya Raipur (IIIT-NR). We thank our colleagues from IIIT-NR who provided insight that greatly helps us in this research. We would like to show our gratitude to Dr. Vivek Tiwari, Asst. Prof. CSE, IIIT-NR for mentoring us and sharing his experience and knowledge with us during this research. We thank every person associated with this research directly or indirectly.

References

1. Li, R., Fu, D., & Zheng, Z. (2017). An Analysis of the Correlation between Internet Public Opinion and Stock Market. In 2017 4th International Conference on Information Science and Control Engineering (ICISCE) (pp. 150–153). IEEE.
2. Tiwari, V., Gupta, S., & Tiwari V. (2010). Association rule mining: A graph based approach for mining frequent itemsets. International Conference of Networking and Information Technology (ICNIT) (pp. 309–313). IEEE.

drksujatha23@gmail.com

3. Kunal, S., Saha, A., Varma, A., & Tiwari, V. (2018). Textual Dissection of Live Twitter Reviews using Naive Bayes. *Procedia Computer Science, 132*, 307–313. Elsevier.
4. Khaidem, L., Saha, S., & Dey, S. R. (2016). Predicting the direction of stock market prices using random forest. arXiv preprint arXiv:1605.00003.
5. Al-Nasseri, A., & Ali, F. M. (2018). What does investors' online divergence of opinion tell us about stock returns and trading volume?. *Journal of Business Research, 86*, 166–178.
6. Lin, Y., Guo, H., & Hu, J. (2013). An SVM-based approach for stock market trend prediction. In Neural Networks (IJCNN), The 2013 International Joint Conference on (pp. 1–7). IEEE.
7. Gupta, A., & Dhingra, B. (2012). Stock market prediction using hidden markov models. In Engineering and Systems (SCES), 2012 Students Conference on (pp. 1–4). IEEE.
8. Bhuriya, D., Kaushal, G., Sharma, A., & Singh, U. (2017). Stock market prediction using linear regression. International conference of electronics, communication and aerospace technology (ICECA), Coimbatore, India. IEEE.
9. Yang, J., Rao, R., Hong, P., & Ding, P. (2016). Ensemble model for stock price movement trend prediction on different investing periods. In Computational Intelligence and Security (CIS), 2016 12th International Conference on (pp. 358–361). IEEE.
10. Labiad, B., Berrado, A., & Benabbou, L. (2016). Machine learning techniques for short term stock movements classification for Moroccan stock exchange. In Intelligent Systems: Theories and Applications (SITA), 2016 11th International Conference on (pp. 1–6). IEEE.
11. Mingyue, Q., Cheng, L., & Yu, S. (2016). Application of the Artifical Neural Network in predicting the direction of stock market index. In 2016 10th International Conference on Complex, Intelligent, and Software Intensive Systems (CISIS) (pp. 219–223). IEEE.
12. Mithani, F., Machchhar, S., & Jasdanwala, F. (2016). A modified bpn approach for stock market prediction. In Computational Intelligence and Computing Research (ICCIC), 2016 IEEE International Conference on (pp. 1–4). IEEE.
13. Sharma, A., Bhuriya, D., & Singh, U. (2017). Survey of stock market prediction using machine learning approach. International conference of electronics, communication and aerospace technology (ICECA), Coimbatore, India. IEEE.
14. George, S., & Changat, M. (2017). Network approach for stock market data mining and portfolio analysis. In Networks & Advances in Computational Technologies (NetACT), 2017 International Conference on (pp. 251–256). IEEE.
15. Dewan, A., & Sharma, M. (2015). Prediction of heart disease using a hybrid technique in data mining classification. In Computing for Sustainable Global Development (INDIACom), 2015 2nd International Conference on (pp. 704–706). IEEE.
16. Sak, H., Senior, A., & Beaufays, F. (2014). Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In Fifteenth annual conference of the international speech communication association.
17. Heaton, J., Polson, N., & Witte, J. (2017). Deep learning for finance: deep portfolios. *Applied Stochastic Models in Business and Industry*, *33*(1).
18. Mrcela, L., Mercep, A., Begusic, S., & Kostanjcar, Z. (2017) Portfolio optimization using preference relation based on statistical arbitrage. In Smart Systems and Technologies (SST), 2017 International Conference on (pp. 161–165). IEEE.
19. Fang, J., & Xiaoyun, M. (2017) Dynamic Multi-Mode Portfolio Optimization Strategy for Markovian Arrival Process. In Robots & Intelligent System (ICRIS), 2017 International Conference on (pp. 139–142). IEEE.
20. Das, S., & Goyal, M. (2012) Rebalancing a two-asset Markowitz portfolio: A fundamental analysis. In Computational Intelligence for Financial Engineering & Economics (CIFE), 2012 IEEE Conference on (pp. 1–8). IEEE.
21. Yang, Y., & Hasuike, T. (2017) Construction of Investor Sentiment Index in the Chinese Stock Market. In Advanced Applied Informatics (IIAI–AAI), 2017 6th IIAI International Congress on (pp. 23–28). IEEE.

drksujatha23@gmail.com

Honeypots and Its Deployment: A Review



Neeraj Bhagat and Bhavna Arora

Abstract Over the last few decades, there has been a tremendous study on the security of networks. The type of data that travels through these networks may contain malicious software, which could harm the systems in the network or affect the network as a whole. This could range from spreading malware to performing active attacks by sending malicious data packets on the networks. However, in the contemporary era of digital world, the information security has become a key area of concern at personal as well as organizational levels. Many methods and tools are used to provide the security in an information system. These methods include intrusion detection and prevention systems, encryption, firewalls, etc. In addition to these, honeypot systems are proposed as complementary structures. Honeypots can be used as traps to attract hackers and attackers in addition to provide support for detection and prevention of intrusions in the system concerned. How honeypots work, its types, and how the deployment of honeypots is done in a network are discussed in detail in this paper.

Keywords Security · Intrusion detection system · Honeypot

1 Introduction

Due to rapid advancement in computer and communications technology, network attack activities have also increased that made computer systems vulnerable to attacks. With the increase in cyberattacks and cybercrimes, the evolution of information security is a primary concern. With every growing year, it is becoming utmost important to enhance the security mechanisms so that the crime and attacks are dealt with. There is an immense need of an efficient system that successfully

N. Bhagat (✉) · B. Arora
Department of Computer Science & IT, Central University of Jammu, Jammu, India
e-mail: neerajb226@gmail.com

B. Arora
e-mail: bhavna.aroramakin@gmail.com

© Springer Nature Singapore Pte Ltd. 2019

V. S. Rathore et al. (eds.), *Emerging Trends in Expert Applications and Security*,
Advances in Intelligent Systems and Computing 841,
https://doi.org/10.1007/978-981-13-2285-3_59

drksujatha23@gmail.com

prevents and detects intrusion on network. Therefore, security of network becomes a great concern for industries that securing the critical information. Some of the most common types of network assault are intrusion of malware, denial of service, eavesdropping, identity spoofing, and password based attacks. To conquest over a lot of these types of assaults, installation of intrusion detection system has to be done that inspects all incoming and outgoing network activity and identifies malware that attack into the system. Honeypot is the fast-emerging technology for the analysis of malicious network traffic. Deployment of a honeypot system on network is proactive measure that enables an immediate detection of an intrusion before any data is damaged or stolen.

This paper is organized in four sections. After the introduction to need of security in Sect. 1, Sect. 2 describes the review of literature of the honeypot technology. Honeypot Technology, its various types and functions are discussed at length in Sect. 3. Section 3 also outlines the deployment of honeypot based on its types along with the tabular representation giving brief description of each form, its pros and cons and efficiency of each of them. Section 4 provides the comparative analysis of honeypot technology. Finally, the conclusion and future work are presented in Sect. 5.

2 Literature Review

The author in paper [1] have implemented a honeypot detection system in their study for detection of malicious web URL. The system that has been served in client side is developed in Python language. By means of crawler, the URL addresses are gathered in the client side and if there is a need to gather more data, corresponding websites are visited. Signature-based intrusion detection system is deployed to check for the malicious content and other vulnerabilities, and if such content is found, a trigger is activated. The URLs containing malicious contents as detected by the system are blacklisted and hence the security is preserved.

Honeypot system for avoidance of malfunctioning in wireless networks is implemented in paper [2]. They have achieved it in phases. In the first phase, they have collected the basic functionality and information about simulator, basic functions of a honeypot, intrusion detection systems, etc. In the second phase, they have created the network with intrusion detection environment in NS2 simulator and fetched the difference in the performance of the wireless network. In third phase, they have configured the honeypot after the firewall to achieve paramount security in the network.

Kumar et al. [3] have used honeypot as a tool to capture unknown and new malware. They have divided their proposed honeypot system into honeypot server and a thin client. About 4 weeks, the author has operated the setup of honeypots and analyzed manually the variations of network traffic which are considered to be malicious by it.

drksujatha23@gmail.com

For the analysis and visualization of malicious activities and connections, the author in [4] has used the honeypot system. Two different honeypots for searching have been set up by the author in their application. Self-propagation option is used as the first searching honeypot and is intended to gather information about malicious software and trap system has been used as the second to gather malicious activities in the system.

A mixed interaction honeypot-based intrusion detection system has been proposed by the author in [5]. They have developed a system to stabilize the network and give details on how the security can be enhanced. Also, the purpose of the system has also been explained in detail by the author.

A distributed honeypot system has been proposed by [6] in order that new vulnerabilities can be searched. In the proposed model, high-interaction honeypots exposure has been increased for the threats coming from the internet by making use of low-interaction honeypots as front-end content filters.

A honeypot-based intrusion detection system has been deployed by the author in paper [7]. It uses IP-based trace back technique for its deployment. The limitation of conventional intrusion detection systems on honeypot systems has been introduced.

3 Honeypot Technology

Honeypots have emerged as a revolutionizing technology in the field of network security and has immense potential to bring new aspects to the security level in organizations as well as personal domains. Diversion of attackers and hackers from critical resources is its major task. Honeypots are considered as observed traps which are used to divert the intruders. Honeypot systems are simulated as real systems on the network but, in fact, they are isolated and monitored closely for the detection of any malicious event. Hence, any user with legitimate actions will never enter the honeypot system and neither have any production value for the system. Only the malicious activities are trapped or are intended to get trapped in the honeypot. Honeypots provide a large amount of valuable information, which is used for analysis and detection of variety of attacks, even in the encrypted environment.

Functions of honeypots

- Divert the attackers away from the real system or network, so that the useful resources are not harmed.
- Intend to consider new types of worms and viruses for future study.
- Profiles of attackers are generated along with their preferred attack methods and modus operandi so that prior information is gathered.
- Identify new types of risks and vulnerabilities of various operating system and also study about behavior of hackers in social network.

drksujatha23@gmail.com

3.1 Types of Honeypot

Honeypots can be classified based on their interaction levels and deployment.

Based on Interaction Levels

i. Low-Interaction Honeypot: In these types of honeypots instead of using the real operating system on which a potential attacker can attack, an emulator is used which interacts with the attacker. Such honeypots offer limited interaction level to attackers. These honeypots are used to scan the port and generate attack signatures (Fig. 1).
ii. High-Interaction Honeypot: In high-interaction honeypots, adversary gains full access to the system and nothing is emulated. Real services are used which motivate the attacker to attack so that attack strategy of attacker can be recorded and analyzed later. These offers 24×7 Internet connectivity to attackers. However, in order to monitor the activities that are performed by the intruders, external programs can be used (Fig. 2).

Based on Deployment

i. Production Honeypots: The implementation of the production honeypots is straightforward as it acquires partial amount of information. To improve the overall efficiency and performance of security, production honeypots are placed

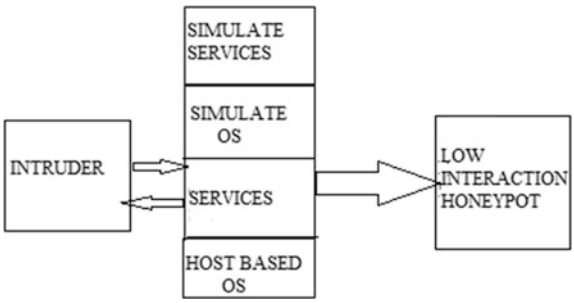


Fig. 1 Low-interaction honeypots [2]

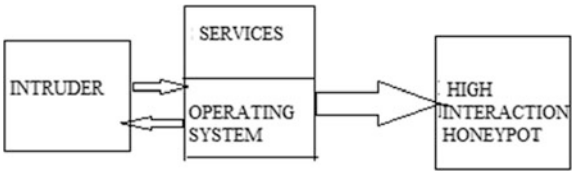


Fig. 2 High-interaction honeypots [2]

drksujatha23@gmail.com

Table 1 Comparison of honeypots based on types

	Based on interaction level		Based on deployment	
	Low interaction	High interaction	Production deployment	Research deployment
Definition	Operating system emulators are installed that interact with the attacker	Real operating system is used	Placed with other production servers inside the network to improve the overall state of security	Run to collect information about the attacker that target different networks
Pros	Simple to use	Add complexity	Easy to deploy	Complex to deploy
Cons	Less prone to infection and risks	High prone to infection and risks	Provide limited amount of information about the attacker	Not append direct value to an organization
Utility	Less utility	More utility	Less utility	More utility
Efficiency	Provide less efficiency	Provide high efficiency	Less efficient	More efficient

with other production servers inside the production network. These are the variation of low-interaction honeypots, which are easier for deployment and maintenance. These provide fewer details about the attacker and the kind of attack than research honeypots.

ii. Research Honeypots: They run to collect information about the attacker that target different networks. Research honeypots provide a platform to research the threats faced by an organization rather than appending direct values. These honeypots are considered complex for deployment and maintenance and extensive information is captured by them. Research honeypots find their suitability in deployment at research centers, military, or government organizations.

Table 1 gives the comparison of these techniques based on its types along with their pros and cons and their utility and efficiency as shown below.

3.2 Deployment of Honeypot

The concept for deploying a honeypot system is to set up a system that appears to be much vulnerable for ease of access to its resources. Honeypots can detect attacks by capturing and analyzing information about variety of attacks, acquiring signatures and working with encrypted data. Honeypots are tools that can be used for surveillance and monitoring of networks. It must be handled with care because it can carry risks to a network due to open access by an attacker (Fig. 3).

drksujatha23@gmail.com

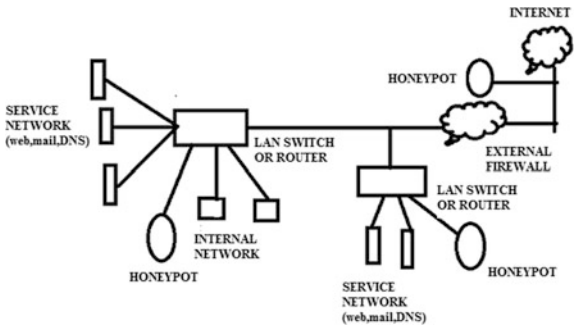


Fig. 3 Deployment of honeypots [8]

Honeypots can be deployed in a variety of locations on a network. A honeypot outside the external firewall is useful for checking attempts to scan attacks on internal network. The main advantages of placing the honeypot at this location are as follows:

- It does not have any side effect that means it does not increase the risk for the internal network.
- Since it attracts and traps attacks to the honeypot, it reduces the amount of traffic, in particular, the attack traffic to the firewall. Therefore, it reduces the number of alerts produced by external firewall.

Honeypot at this location does not trap internal attackers. Honeypot can also be placed in the demilitarized zone (DMZ) to anticipate threat warnings that can emerge of the network. Public domain servers that are accessible to users freely are effortlessly imitated by the honeypot system. Production environment security is also increased because of the inadequate access to internal network from the DMZ. The main disadvantage of placing honeypot at this location is that DMZ is not fully accessible.

Honeypots can also be placed in the internal network along with some service network stations. The main advantages of placing the honeypot at this location are:

- It can catch internal attacks.
- It can also detect a misconfigured firewall.
- On the other hand, unless we can completely trap the attacker within the honeypot, the attack may be able to reach other internal systems from the honeypot. In order to continue to attack and trap the attacker to the honeypot, we must allow this attack traffic from inner network to their honeypot. This means that we must open up the firewall to allow the attack traffic to come from the internal network and this carry the huge security risk. The firewall must adjust its filtering to allow traffic to the honeypot.

drksujatha23@gmail.com

4 Comparative Analysis

This section presents a brief summary and comparison of work that has been done by researchers in this field. It can be seen that the work based on honeypot technology for intrusion detection is gaining a wider acceptance in organizations to achieve security. Table 2 presents a brief comparison of honeypot technology by using different tools and algorithms that have been used to set up a honeypot system.

Table 2 Comparative analysis of work using honeypots

	Kumar et al. [3]	Shukla et al. [1]	Chawda et al. [6]	Koniaris et al. [4]	Joshi et al. [2]
Year	2009	2014	2014	2014	2017
Tools/ Algorithms used	Bloom filter	JavaScript tool (SpiderMonkey), KaliLinux, Python API Beautiful soup	Honeyd, POf, Snort, Nmap	Dionaea, MySql, Honeyd, QGoogle	NS2 simulator
Inputs	Different window services (PCS, CRS, MDS)	Webpage with 100 malicious URL link	Database containing host description and log information.	Malicious connections and IP address taken	Wireless network
Analysis	Analysis of type of malicious network traffic	Analysis of blacklisted files which contains a list of IP addresses triggering vulnerabilities of different OS	Analysis of dynamic and hybrid model for tracking and characterizing internet threats such as worms or automated attacks	Analysis and visualization of malicious activity and connections	Analysis of performance of honeypot-based IDS by reducing energy spent and packet drop rate
Outputs	Detected malicious strings consisted of HTML code and Java Script	Total detected events–27 malicious URL detected–6 Unique event detected–11	Provided fine grained output about the threat	Observed connections per protocol/ per day/per week/unique IP address	Packet drop rate reduced

drksujatha23@gmail.com

5 Conclusion and Future Work

Like all technologies, honeypots have their pros and cons. Honeypot systems are seen as a complimentary technology to network security and host-based intrusion protection system by the security experts and they do not recommend these systems as replacement for the existing security technologies. This paper represents basic technology of honeypot and their deployment. Conventionally, honeypots have been deployed against external or common internal threats. However, the research in this area is still in the early stages and many open research challenges remain unresolved and can be explored and implemented in various network frameworks for attack detection.

References

1. Shukla R (2014) PythonHoneyMonkey: detecting malicious web URLs on client side honeypot systems, pp 0–4
2. Joshi V, Kakkar P (2017) Honeypot based intrusion detection system with snooping agents and hash tags, vol 8(2), pp 237–242
3. Kumar S, Pant D (2009) Detection and prevention of new and unknown malware using honeypots. Int J Comput Sci Eng 1(2):56–61
4. Koniaris I, Papadimitriou G, Nicopolitidis P, Obaidat M, Ieee F (2014) Honeypots deployment for the analysis and visualization of malware activity and malicious connections, pp 1819–1824
5. Zou Q A new type of intrusion prevention system, pp 0–3
6. Chawda K (2014) Dynamic & hybrid honeypot model for scalable network monitoring, no 978
7. Dongxia L (2012) An intrusion detection system based on honeypot technology, pp 451–454
8. https://en.ppt-online.org/79721

drksujatha23@gmail.com

Study on Data Mining with Drug Discovery



Bahul Diwan and Shweta Bhardwaj

Abstract An important goal of our health system is to classify new drug events that are ADE in the post-approval period. Data mining methodologies that can transform the data set into meaningful knowledge informing patient safety has proved essential for this process. This research paper describes the application of the biomedical documents to facilitate knowledge from the very large data set that comprises of drugs and the characteristics of Life Science. This paper helps in discovering the data or discovering the pattern from the large data set and demonstrate the technique of data mining that can help in the quality of the decision-making process in the medical industry.

Keywords Data mining · Drug discovery · Drug · Clustering
Dataset · Pharmacy

1 Introduction

Almost three decades ago, information in the medical industry was relatively simple and application of the technology was not that complex [1], however, as we have moved forward, technology became complex but also its application usage has become easy to implement.

Data is the new gold. Data mining is used for extracting information from the larger dataset using different data structures, algorithms, and techniques in the field of mathematics and statistics. Database management system, machine learning, and traditional data analysis method required more time and analyzing data manually was expensive. It has been predicted that the revenue growth from the medical

B. Diwan · S. Bhardwaj (✉)
Computer Science and Engineering Department, Amity University Uttar Pradesh, Noida, Uttar Pradesh, India
e-mail: shwetabhardwaj84@gmail.com

B. Diwan
e-mail: bahul1205@gmail.com

© Springer Nature Singapore Pte Ltd. 2019

V. S. Rathore et al. (eds.), *Emerging Trends in Expert Applications and Security*,
Advances in Intelligent Systems and Computing 841,
https://doi.org/10.1007/978-981-13-2285-3_60

drksujatha23@gmail.com

industry will slow down from 12.1% to 5.1–6.1% rate. In drug discovery, we can use chemoinformatics, the use of computer [2] and IT field that is being applied to a range of problems in the field of chemistry and we can use the same methodology for the drug discovery as it transforms data into the useful format that we can use for the decisions in the drug field. For example, a data mining contest was being held to predict the bioactivity for the drug design, specifically data mining in which molecule that is being organic would bind to a target site on thrombin. The futuristic prediction based on about 450 MB of data on approximately 2000 organic molecules each with more than 130000 attributes, was a challenge in itself not only because of the large number of attributes, but because only 43 of the compounds were active, only 10% were able to achieve the results with more than 61% accuracy.

In this paper, we focus on how data mining discovers and extracts the useful pattern from the larger dataset so that we are able to find the observable patterns. The importance of data mining is improving the quality of future or the quality of the decision-making process in the medical industry.

2 Data Mining Technique

Medical industry depends on decision making, systematic selection of the models that enable the decision evaluator to evaluate the payoff that has been accepted as a result of the implementation of the selected program. There are many different data mining techniques that can be applied on the data warehouse to obtain the useful information and our aim is to study, analyze various mining techniques, and apply those algorithms in the real time [3]. We have to use chemoinformatics and the mining technique to find the unknown drug that is similar to a certain specific known drug.

Six important techniques that are being used in data mining processes are as follows:

- Problems defined
- Knowledge acquired
- Data selected
- Data preprocessing
- Analysis and interpreting
- Report generation.

The techniques used in data mining are as follows:

A. **Association**

This method is used to identify the rule of affinity among the collection or the collected set. The application of the association rule is that it includes market basket analysis, mailing in the direct market, fraud check [4].

drksujatha23@gmail.com

B. **Prediction and classification**

Prediction and classification model are two data analytic techniques that are being used to tell about the classes of the data and predict the futuristic data classes [5]. Debit Card Company can classify its customer dataset as per poor, medium or good. Similarly, the income level of the customer can be classified as low, medium, and high. If we have a record containing the behavior of the customer and we want to classify the data or make a prediction, we find the task of prediction and classification as they are very closely linked. Model of the neural networks decision tree is based on classification schemes and they are very much useful in the medical industry.

Classification works on the discrete and unordered data is described in [6] while regression is often used as the statistical method used for numeric prediction. The primary focus should be made on the selection of measuring accuracy and predictive efficiency of discovery of a drug.

C. **Clustering**

Clustering is the method in which similar records are being grouped together to form clusters. An organization can take the hierarchy of the classes that are being grouped for the similar events. Using the method of the class, string employee can be grouped on the basis of their age, income, occupation, housing, etc. In business, clustering helps to identify a group of class on the basis of similarity and help them to form clusters on the basis of their purchasing pattern etc.

3 Relevance of Data Mining in Drug Discovery

Different mining techniques can be used on the warehouse of the data to obtain the knowledge or the information that is useful [7]. Our work is to analyze and study the various mining techniques and apply those in the real world using various algorithms. The application named drug discovery is being developed by keeping in mind the slow process of developing a new drug. Drug discovery in today's era is used to conceptualize the involvement of the chemoinformatics to overcome the shortcoming of the previous drug development processes that are slow. The platform that is being used for creating the drug application is primarily in Java. NetBeans is used and use Microsoft Access in the backend for storing the drug data sets.

The system allows the user to find the unknown cluster of the similar drugs and help them to compare with the different clusters example HIV, cancer drugs [8]. It helps us to know who discovered the new drug from the unknown drug that is similar in property with the known drug.

Starts inference is an assumption driven in the sense that a hypothesis is formed and tested against data mining based on discovery driven and is robust for real-world data.

drksujatha23@gmail.com

Mining can answer analytical questions such as what is the discovery of new molecules and what are the issues over it? What factor or what combination is directly impacting the drugs? How to optimize or to allocate resources to ensure effectiveness and efficiency?

4 Applications of Data Mining in the Medical Industry

Most of the healthcare Institutes use information system as they cannot produce reliable reports with respect to the information. The management of the medical industry starts to recognize the relevance of the drugs and their definitions. In the connection between the care result and cost and the patient satisfaction, the right technology is needed and can be found using the information and technology that is being used as a communication process.

The delivery of the health care has always been related to the information and the industry is recognizing the increase in the use of automated systems and thus focuses on production. Operational data and nonoperational data are being used in the automated systems. These systems are being named as legacy systems to be used for administration purpose but also for the data process purpose. Plot information is been hidden in the system, that is being named as legacy because there are no answers to the questions that are unpredictable.

User interface can be built or designed that can accept the information from the user regarding age, food consumed, etc. This data can be entered into a database and the relevant tool can be used to extract the information regarding that patient or a user. The profile needs to be anonymous as it contains some vital information about the user.

One of the major problems with medical data is actually lack of information. For example, Drug and Food Administration Department estimated that only about 2% of serious events are reported. Food and drug fear of litigation may be a contributing factor in this and lack of healthcare providers that do not fill the reports as they find it as a time-consuming process.

5 Implementation of Data Mining for Drug Development

Data Preprocessing: Used for the preprocessing step.

Data Selection: This process involves selecting of data marts that are used in the discovery. The data used in the KDD process is selected based on the evaluation of its field knowledge, the data source is individually referred to as data marts.

Data Integration: At this stage, multiple sources of data are combined in a common source called data warehouse that is being used to centralize data.

drksujatha23@gmail.com

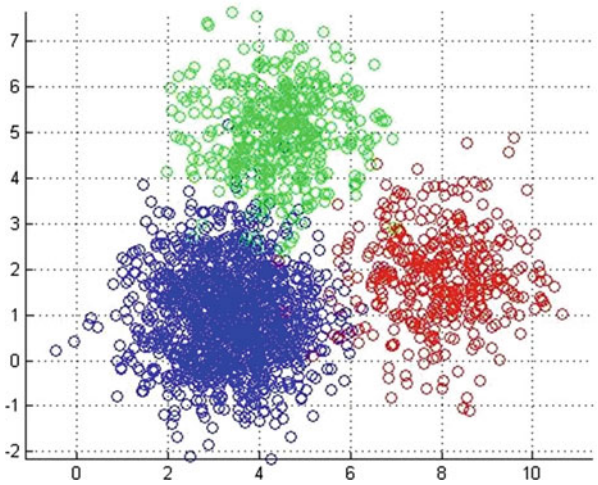


Fig. 1 K-mean clustering [1]

Characteristics employed to the data that is being stored in a data warehouse are time independent, nonvolatile, and subject oriented (Fig. 1).

The integration process is as follows:

- N data marts
- Create a database
- Each data mart (1 to n)
- For each compound in the data mart
- If the compound name does not already exist in the database
- Add compound to database
- Else
- Go to next compound
- End if
- End inner loop
- End outer loop.

Data Cleaning: Data cleaning is intense and involves the examination of the data for its completeness and integrity.

The cleaning process involves following steps:

- For each structure in the database
- Check the missing field
- If found
- Take a default value from the user or assign null
- End if
- End for loop.

drksujatha23@gmail.com

Data Mining: Component of the data mining use K means clustering technique. K mean clustering is a method of a cluster analysis, which aims to partition and observe the K cluster in which each observation belongs to a cluster with the nearest mean [5]. The main idea of this technique is used to define the centroid for each cluster. We have to place each cluster as far away as we can and then we take a point that belongs to a dataset and associate it to the near centroids. When there is no point pending, the first step is completed and the early age group is done. Now we have to calculate new k centroids and they become the bar center of the cluster resulting from the previous step. We have to keep doing this loop till the time the centroid does not move anymore (Fig. 2).

The algorithm is composed of the following steps:

- Place the points into the space that is being represented as objects and part of the cluster. These are the initial centroids.
- Assign each specific class or object to a group that has the closest centroid when all objects have been assigned to calculate the k centroid.
- Repeat the above steps till the time the centroid does not move anymore.

Pattern evaluation: In this step, the cluster of the unknown similar drugs are being evaluated and compared with the cluster of some of the specific drugs that are been already discovered. Representing horror search in a visual way is the new feature of the application. The second step in this process is the calculation of the structural similarity between the drugs using the known Tanimoto coefficient.

Smile: Smile is a simplified molecular input line entry system [9]. It is a notation for entering and representing chemical compound using the ASCII. We use smile notation to store the structure of a chemical compound.

Fingerprints: Fingerprints are a very abstract representation of a structural feature of a molecule, in a form of a string consisting of 0 and 1. We use Smile notations for the fingerprint process.

Following algorithm is being used for the fingerprint process.

- For each compound in the database
- Fingerprint = 0
- For each substructure Sub in a database

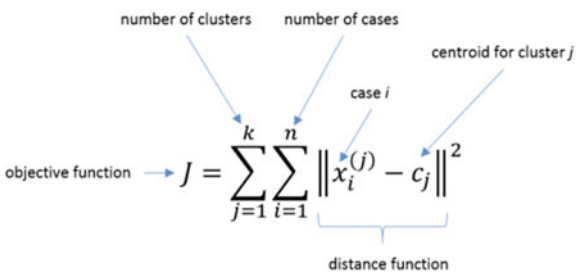


Fig. 2 Formula used [10]

drksujatha23@gmail.com

- If Sub is subpart of the chemical compound
- Fingerprint = fingerprint + 1
- Else
- Fingerprint = fingerprint + 0
- End of inner loop
- End of outer loop.

Tanimoto Coefficients: This term is used to find the similarity between the two chemical compounds and the basis that is being required for this process is the fingerprint of that compound.

1 denotes subpart, 0 denotes the absence of substructure of the compound.

Tanimoto coefficients between two compounds A and B

$$T = NAB/NA + NB - NAB$$

where NAB denotes the number of bits that are 1 in the fingerprint of both the compounds.

NA denotes the number of bits that are 1 in compound A.

NB denotes the number of bits that are 1 in compound B.

6 Conclusion

About 61% (118) drugs of 201 drugs form a cluster having the molecular weight (74.9–408); CLOGP (−6.3–6.1); H.ACC (1–10); H.DON (0–5) and about 6.99% drugs out of 201 drugs form clusters having the molecular weight (111–408); ClOGP (−4.0–4.3); H.ACC (2–8); H.DON (0–3). There are many other small clusters.

7 Result

Using the gdsc dataset, genomics of drug sensitivity center focus on the drug response to the cell data line. We focus on two parts—sensitivity and resistivity. With the help of previous data, we understand the functionality of it. We create two things—cell network and resistance network. We focus on checking the structure that is similar to one another and then calculates drug response.

In Fig. 3, ($y = x$) is the basic classification lines of the resistant and sensitive pair as any point under the line might have a high sensitive score than resistance.

drksujatha23@gmail.com

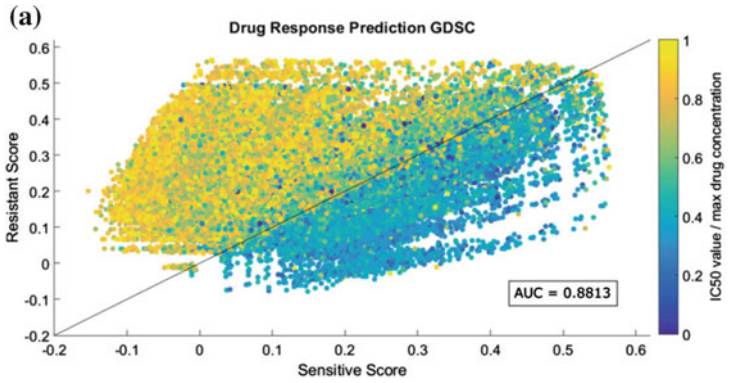


Fig. 3 Result

Link predictive analysis

- **check the sensitivity score**
- **check the resistivity score**
- **check the structure**
- **place according to line**
- **y axis = resistance**
- **x axis = sensitivity**
- **yellow = resistivity**
- **blue = sensitivity.**

References

1. http://blog.thedigitalgroup.com/rajendras/2015/05/15/k-means-clustering/
2. http://pubs.acs.org/doi/abs/10.1021/ci025620t
3. http://www.bmj.com/content/322/7296/1207.short
4. https://academic.oup.com/nar/article/34/suppl_1/D668/1132926
5. http://searchsqlserver.techtarget.com/definition/data-mining
6. Han J, Kamber M Data mining concepts and technology
7. https://www.cambridge.org/core/journals/knowledge-engineering-review/article/a-survey-of-knowledge-discovery-and-data-mining-process-models/368D6AFE435EB5E30378398D34D61C17#fndtn-metrics
8. http://www.cancerrxgene.org/
9. https://images.nature.com/original/nature-assets/srep/2017/170109/srep40321/extref/srep40321-s1.pdf
10. https://home.deib.polimi.it/matteucc/Clustering/tutorial_html/kmeans.html

drksujatha23@gmail.com

Efficient Hybrid Recommendation Model Based on Content and Collaborative Filtering Approach



Ankita Gupta, Alok Barddhan, Nidhi Jain and Praveen Kumar

Abstract Recommendation systems are employed in e-commerce and market analysis applications and websites adaptable to customer requirements and interests. These systems analyses trends and people preferences and promote market strategies to enhance businesses. Such recommendation system is built purely with the science of understanding large sets of data generated and collected from the people and can be used to mobilize market trends. In this paper, a novel architectural model for recommendation systems has been proposed. The approach aims at overcoming the limitations of the traditional recommendation Systems. A hybrid of content-based filtering and collaborative based filtering techniques are proposed that spans different item-user parameters for making recommendations. The similarity indices are computed using various mathematical models like cosine similarity, centered cosine similarity etc.

Keywords Hybrid recommendation system · Content-based filtering
Collaborative filtering · Rating normalization · Matrix factorization

Please note that the LNCS Editorial assumes that all authors have used the Western naming convention, with given names preceding surnames. This determines the structure of the names in the running heads and the author index.

A. Gupta (✉) · P. Kumar
Department of Computer Science and Engineering, Amity School of Engineering and Technology, Noida, Uttar Pradesh, India
e-mail: ankita.er.gupta@gmail.com

A. Barddhan · N. Jain
Centre for Development in Advanced Computing, Noida, Uttar Pradesh, India

© Springer Nature Singapore Pte Ltd. 2019

V. S. Rathore et al. (eds.), *Emerging Trends in Expert Applications and Security*,
Advances in Intelligent Systems and Computing 841,
https://doi.org/10.1007/978-981-13-2285-3_61

drksujatha23@gmail.com

1 Introduction

Recommendation Systems form a major part of the digital world, where technology enables individual, easy access to a variety of substitutes and varieties of commodities and services at a finger tap. The various filtering techniques can be developed into recommendation engines that provide the intelligence to applications to provide personalized and user-friendly interaction with individuals in terms of their preferences and choices. In this research work, a novel hybrid recommendation approach has been developed keeping in mind the various factors that can affect the user preferences in music. The proposed model tries to overcome the combined limitations of content based and collaborative filtering approaches and provide fast and efficient solutions to the recommendation problems [1].

The existing recommender systems used so far generates a recommendation for the user in three ways: Content-based Approach, Collaborative Approach, and a Hybrid approach that is a combination of more than one recommendation approaches [2].

2 Methodology

2.1 Parameters

The recommendations are trained from a set of a database of the music library and user access data. The parameters used for recommendations are—Song Title, Song Name (Version based), List of Artist Names, User ID, Play count of Song per User.

2.2 Dataset

The tables consist of information about all the songs consisting of 1,000,000 record sets (Table 1).

2.3 Proposed Model

The proposed model for music recommendation implements the listed flow of steps in providing a hybrid recommendation. See Fig. 1.

1. Add a song name from the current user.
2. Compute content-based similarity on the basis of artist between the entered song and database of all other songs relevant for the recommendation.

drksujatha23@gmail.com

Table 1 Dataset

Attribute	Description
Track ID	A unique ID assigned to every song that acts as a primary data field for the entire dataset for identification of the song through different relations and tables
Song ID	The Song ID is a unique ID given to every track in the collection of the dataset. note that track ID assigned to a particular song is the same, whereas different versions and renditions of the song (for e.g., unplugged covers, acoustic cover etc.) are given different Song ID's. hence song ID is the most unique ID to define a song as an independent unit
Song artist	This attribute contains the names of the artists involved in the composition. There can be multiple artists separated by special characters
Song title	This field consists of the name of the song
User ID	Every user is assigned with a user ID that identifies each user uniquely
Play count	Play count is the total number of times the song played by the user reflecting his affinity to the song

3. If other users exist from the user table, who have listened to the current song, then proceed to next step or else generate the list of recommendations based on content-based similarity.
4. If the rating data is available from the user table then proceed for collaborative filtering similarity, else generate ratings through prediction.
5. After collaborative filtering, release the list of top-n recommendations to the user.

3 Implementation

The proposed model is implemented using R scripts. Data frames and vectors are used to store intermediary computational results and processing arithmetic processing.

3.1 Content-Based Similarity Computation

For the computation of content based similarity between the user entered song and the collection of all other songs present in the database, we use artist as a common factor for similarity measure between songs. The play count and the number of artists act as two vectors for which Cosine Similarity measure can be calculated [3, 4]. Cosine similarity formulae can be modified as follows:

drksujatha23@gmail.com

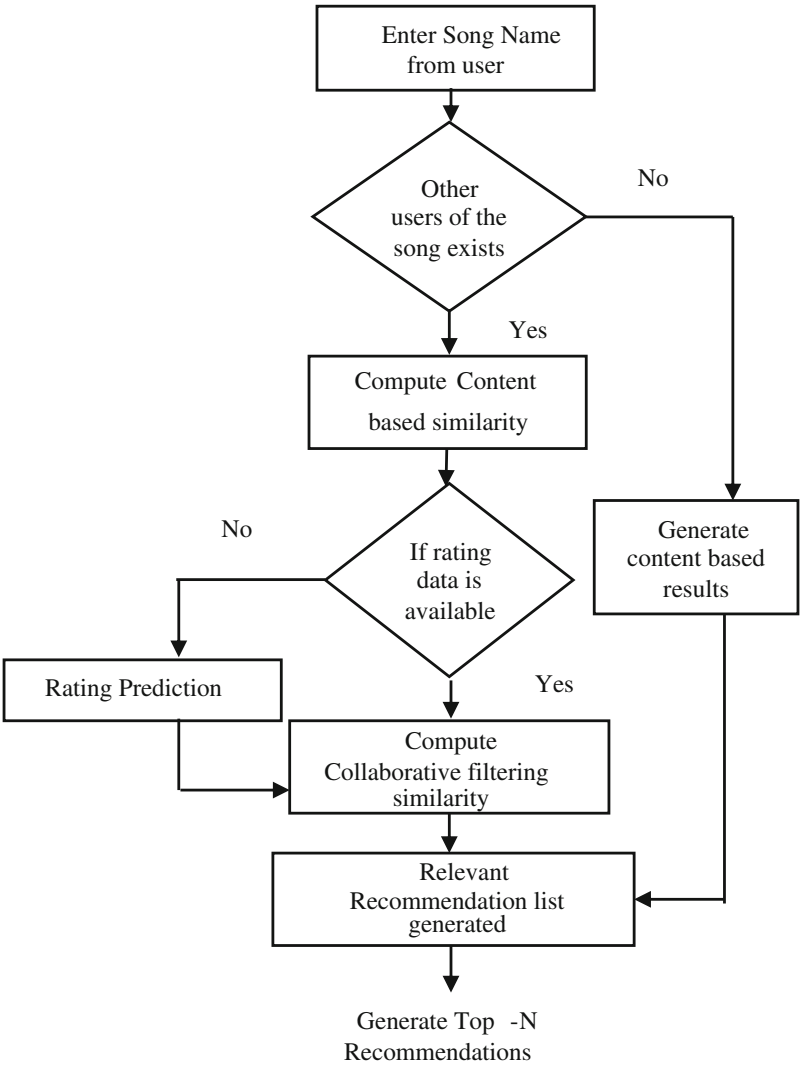


Fig. 1 Proposed model

$$\text{Cosine Similarity} = \frac{\text{dot product}}{\text{length of the vectors}}$$

$$= \frac{(\text{No. of Common Artists to both songs}) + (\text{Product of average rating})}{\left(\sqrt{\textbf{No. of artists for song } 1 + (\textbf{Rating of Song } 1)^2}\right) * \sqrt{\textbf{No. of artists for song } 2 + (\textbf{Rating of Song } 2)^2}}$$

drksujatha23@gmail.com

Table 2 Content based similarity computation

Songs	Artist names									Count based rating
	a1	a2	a3	a4	a5	a6	a7	a8	–	
Song ID 1	0	1	1	0	1	0	1	0	1	3
Song ID 2	1	1	0	1	0	1	1	1	0	4
Song ID	–	–	–	–	–	–	–	–	–	–

The above calculation is applied to the dataset as follows:

Consider the song 1 as song entered by the user while Song 2 as an item from the database for similarity computation. For every artist that is an attribute of the song, the table is filled with 1, while every artist who has not sung the song, the table is filed with Boolean 0 entries [5]. The count-based rating is obtained by merging the user and song table to obtain the number of play count for each song (Table 2).

1. The dot product of the two vectors is 2 + (3 * 4) while the length product of the two vectors is $(\sqrt{5} + 3 * 3) \times (\sqrt{5} + 4 * 4)$.
2. Thus, the cosine value between the two vectors is computed to be 0.816.

The top-n songs with the highest similarity scores can be further fed as input to the collaborative approach.

3.2 Collaborative Filtering Based Similarity Computation

The content-based similarity computation is executed on a subset of songs that are shortlisted by the feature of collaborative-based filtering. This overrules the exclusivity of the two techniques. The songs that have also been heard by the set of users that have heard by song entered by the current user are shortlisted and the content filtering is executed on this sublist [6].

Collaborative filtering approach uses Centered Cosine approach of computing similarities between the ratings provided by the different users to the 2 songs under comparison [7]. The centered cosine approach is preferred with a rating scheme for the following reasons.

1. It provides centering of the rating scheme around 0, by eliminating negative values that may be generated in case of unavailability of rating for a particular user and song.
2. It captures the intuition of the rating scheme better as missing rating are treated with average values and “tough rates” and “easy rates” are resolved easily (Table 3).

drksujatha23@gmail.com

Table 3 Collaborative filtering-based similarity computation

Users	Song names							Total play count
	S1	S2	S3	S4	S5	S6	S7	
User ID 1	0	20	4	8	2	0	1	35
User ID 2	1	15	0	4	8	2	1	31
User ID	–	–	–	–	–	–	–	

Star Rating Prediction

The rating prediction from play count is carried out in the following manner:

1. The rating is related to the play count of each song by the particular user. Hence we find out the total number of plays for each song by computing the row-wise sum of the entire table.
2. Every element of the cell is divided by the total play value of the corresponding song.
3. The value generated ranges from 0 to 1. Hence it is converted into the scale of 0–5.

Normalization/Centeredness of Rating Scheme

The ratings generated by above schemes 0 ratings for the songs that have not been heard by the users at all. In such a case the overall average rating of the song is affected which should not be the case to analyze the user preferences efficiently. Hence, we normalize the rating scheme by applying the following manner [8].

The average rating of each song is computed and subtracted from each rating value of that corresponding song. Thus, all values are converted to positives for good ratings and negatives for those who have rated them poorly. The computed value is used to find similarity based on a collaborative approach.

Collaborative Similarity Computation

The similarity is computed using cosine formulae from the predicted rating table with the formulae below [9].

$$\text{Cosine Similarity} = \frac{\sum(\textit{Rating for Song 1 by user } i) * (\textit{Rating for Song 2 by user } i)}{\left[\sqrt{\sum(\textit{Rating for song 1 by user } i)^2}\right] * \left[\sqrt{\sum(\textit{Rating for song 2 by user } i)^2}\right]}$$

The calculated similarity is used to order the song for more likeliness to be preferred by the user. And, the top 10 results are displayed to the user.

drksujatha23@gmail.com

4 Future Work

As a future scope of this system, various attributes combined such as genre, Measure (bar), tempo, etc., can be incorporated in this model. Similarly, user attributes like age, taste, location, etc., can be used to refine collaborative filtering techniques.

References

1. Shrote KR, Deorankar AV Review based service recommendation for big data. In: International conference on advances in electrical, electronics, information, communication and bio-informatics (AEEICB16)
2. Wang C, Zheng Z, Yang Z (2014) The research of recommendation system based on hadoop cloud platform. In: 9th of international conference on computer science & education (ICCSE 2014). IEEE
3. Meng S, Tao X, Dou W (2013) A preference-aware service recommendation method on map-reduce. In: 16th international conference on computational science and engineering. IEEE
4. Parvatikar S, Joshi B (2015) Online book recommendation system by using collaborating filtering and association mining. In: International computational intelligence and computing research. IEEE
5. Verma JP, Patel B, Patel A (2015) Big data analysis: recommendation system with hadoop framework. In: International conference on computational intelligence & communication technology. IEEE
6. Kupisz B, Unold O (2015) Collaborative filtering recommendation algorithm based on hadoop and spark. In: International conference on industrial technology. IEEE
7. Ghuli P, Ghosh A, Shettar R (2014) A collaborative filtering recommendation engine in a distributed environment. In: 2014 international conference on contemporary computing and informatics (IC3I)
8. Meng S, Dou W, Zhang X, Chen J (2014) KASR: a keyword-aware service recommendation method on mapreduce for big data applications. IEEE Trans Parallel Distrib Sys
9. He B, Zhang H (2016) Library personalized information recommendation of big data. In: International conference of online analysis and computing science. IEEE

drksujatha23@gmail.com

Research Review on Digital Image Steganography Which Resists Against Compression



Darshan M. Mehta and Dharmendra G. Bhatti

Abstract Image steganography is one attempt toward the conversion of communication between sender and receiver. Most of the steganographic methods and algorithms belong to either spatial or transform domain. Most of the time, secret data embedded in stego image gets lost when stego image gets compressed. This secret data loss problem mainly occurs when stego image gets compressed with a lossy compression scheme with a higher compression ratio. In this research review paper, we have presented a survey of image steganography with focus is in the direction on Digital Image Steganography Which Resists against Compression with parallel consideration of payload, imperceptibility, compression ratio, and performance parameters, which represent research work of many researchers in this area and direction. The paper also includes findings derived from a survey of image steganography-related research work done by many researchers. This paper also gives the conclusion and future scope for further research in digital image steganography.

Keywords Spatial domain · Transform domain · Compression ratio
Payload · Imperceptibility · PSNR · MSE

1 Introduction

Usage of the Internet becomes very vast. People utilize the Internet for information exchange. Users are passing information through Internet, which is a public network. So the first challenge comes up is information security. One attempt toward information security is steganography [1], among others (text, audio, video) one is image

D. M. Mehta (✉)
UCCC & SPBCBA & SDHG College of BCA & IT, Udhna, Surat, Gujarat, India
e-mail: dmmehta83@gmail.com

D. G. Bhatti
Computer Science Department, Uka Tarsadia University, Bardoli, Surat, Gujarat, India
e-mail: dgbhatti@utu.ac.in

© Springer Nature Singapore Pte Ltd. 2019

V. S. Rathore et al. (eds.), *Emerging Trends in Expert Applications and Security*,
Advances in Intelligent Systems and Computing 841,
https://doi.org/10.1007/978-981-13-2285-3_62

drksujatha23@gmail.com

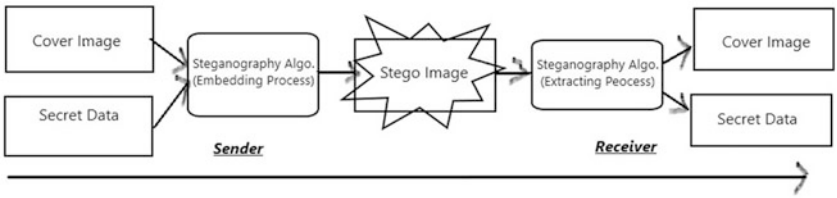


Fig. 1 Basic diagram of image steganography

steganography [2]. Good amount of work has been done about image steganography but when the image gets compressed or distorted, the information within the image is lost [3, 4]. Here, the second challenge comes up that is secret data loss at the receiver end. Another three traditional challenges combined still need to be considered in the future, that is, compression ratio, imperceptibility and higher payload [2, 3, 5–7]. Steganography is the term used to refer to the conversion of communication between entities. Image steganography refers to conceal secret data into an image. The goal of steganography is to hide communication between Entities; it takes less attention of intruders, so the chance of attacks on secret data becomes less. The standard terms used with steganography are: cover image, secret message, and stego image [2]. The cover image is used as a carrier of secret message, the result of steganography algorithm is stego image which contains secret message within it. Figures 1 [2, 8–10] shows the basic diagram of image steganography.

2 Literature Review and Related Work

Bajpai et al. [1] evaluate some of the steganography algorithms operate in the spatial domain and propose a guiding element that resizing the image according to the length of the secret message reduces computation time and enables optimal embedding and transmission over the network. A good approach to steganography must provide high security against compression. There are three main issues to be considered when studying steganography systems: capacity, imperceptibility, and robustness [11]. Robustness refers in the context of image steganography means survival against the possibility of modifying or destroying the secret information embedded inside the image. Steganography techniques using spatial and transform domain are common in use [12]. Both domains have their advantages as spatial domain gives better capacity and transform gives more robustness than the spatial domain. So both capacity and robustness can be achieved by using both domains in a single hybrid method [9]. As per paper [13] two-dimensional DWT is used for steganography in 24-bit color images which is resistant against lossy compression is acceptable. Atawneh et al. [8] proposed a new blind image steganographic method which utilizes both domains combined. The proposed method achieves good performance in terms of image quality, embedding capacity and robust against the lossy compression. In this paper, [2]

drksujatha23@gmail.com

discusses spatial domain techniques having a large capacity, but not able to withstand lossy compression. Whereas DCT-based steganography techniques trade-off robustness with capacity. DWT steganography provides better results compared to DCT. Nowadays, with the extensive research already going on is based on a hybrid domain, or is adaptive. Tamanna et al. [14] have analyzed that no technique of steganography, which provides a higher payload, imperceptibility, and robustness properties together. It is very difficult to achieve high robustness and payload capacity at the same time. Among many attacks, one is compression attack which compresses the image might result in the loss of the secret information embedded in the image. The paper [15] presented a comparative analysis of image compression is done by two transform methods, DWT and IWT. The stego image is derived by applying various combinations of DWT and IWT on both images. The proposed method provides a good amount of payload to cover images with imperceptions and robustness [15]. Furqan et al. [16] implement the digital watermarking algorithm by combining DWT and SVD techniques. The watermarked image gets compressed lossy compression attack than also they succeed to extract the originally inserted watermark image. They suggest that the implemented algorithm can be improved and further can be extended to color images. Kaur et al. [5] proposed an adaptive steganography technique to provide a better trade-off between capacity and robustness. Watermark is a sub-branch of a robust steganography technique. The analysis is performed using PSNR, RMSE with embedding capacity on some benchmark images like Lena, Baboon, etc., for checking the accuracy of the proposed Algorithm [5]. Adwan et al. [17] discussed that DCT and DWT are methods, which embed the secret information within the image in particular portion which is less exposed to compression, cropping, etc. Zhang et al. [3, 6] proposed a JPEG compression-resistant adaptive steganography algorithm based on dither modulation and the relative relationship between DCT coefficients. They intense future direction with the main focus is on the under higher payloads conditions [3, 6]. Das et al. [18] proposed a DCT-based watermarking system. Even after applying high JPEG compression ratio, the embedded watermark can be extracted. Resistance to JPEG compression is very important in this assessment of robustness. Kumar et al. [19] review of different compression of images with their pros and cons. EZW, SPIHT, and Modified SPIHT algorithms are the some of the important compression techniques [19]. Most generalized technique for image compression is the wavelet approach. Mali et al. [10] present a robust and secure method of embedding a high volume of text information in digital cover images without incurring any perceptual distortion. It is robust against image compression attack. Image-Adaptive Energy Thresholding (AET) is used while selecting the embedding locations in the frequency domain. Coding framework with Class-Dependent Coding Scheme (CDCS) along with redundancy and interleaving of embedded information gives enhancement in data hiding capacity [10]. Aulakh et al. [7] presented a comparative study of image compression approaches is done by transform methods like DCT, DWT, and Hybrid (DCT-DWT) transforms. They can obtain higher compression ratio using a hybrid approach, but chances of losing secret information are

drksujatha23@gmail.com

more, DWT gives a better compression ratio without losing more data of image whereas DCT gives less compression ratio but data loss is very negligible [7]. There are many researchers who proposed steganography, but most of their techniques cannot tolerate the destruction of lossy compression. Hwang et al. [20] proposed lossy compression-tolerant data hiding technique is based on the vector quantization and the lossy compression technique of JPEG. It provides a large compression ratio. Compression ratio and image quality are two important features in our proposed data hiding method. However, if we want to get a larger compression ratio of image, the image quality will be bad.

3 Research Methodology

We collect research and review papers on image steganography, after than narrowing our search for reviewing papers which discussed about image steganography survival against compression. We did experiments on some most popular spatial domain methods with some compression methods. We continue Lit. Rev. in the said direction and accordingly we derive our findings and conclusion with a future scope.

4 Findings

We have read papers mentioned in the literature review and related work section of this paper and identified the following findings from above extensive work carried out by many researchers. All the researchers used research methods in their work that either belong to spatial or transform domain. We have tested some algorithms like LSB, PVD, RPS with compression algorithms like RLE, DCT in MATLAB R2012a and result derived that when stego image gets compressed, information within the image is lost entirely, partially, or get ambiguous [2, 9, 14, 19, 21]. Robustness, imperceptibility, cr, secret data size, PSNR, MSE are very important parameters to test the performance of steganography system [21]. Image steganography's resistance to lossy compression with higher payload should be the focus in future [3]. There is a need to design and implement algorithms, which not only provide a wider capacity for the confidential information to be embedded but also remain undetectable and resist lossy compression entirely [7]. Compression-resistant adaptive steganography under higher payloads with improves the correct extraction rates and detection resistant ability of the stego images will be desired in future [22].

drksujatha23@gmail.com

5 Conclusion

Spatial domain methods are very less robust against lossy compression but can achieve higher payload. Transform domain methods prove more robust against lossy compression with the low payload. There is a trade-off between robustness and payload features. Some researchers worked on the hybrid domain or adaptive methods to reduce the trade-off between said features. The above research reviews and findings presented motivates to design and develop an image steganography algorithm development for securing secret information which resists against lossy compression with higher payload, imperceptibility, security together with keeps main focus on lossy compression resistance feature.

6 Future Scope

Designing and developing a digital image steganography algorithm on color image for securing secret information, which resists against lossy compression schemes with achieving higher compression ratio, extraction rate, higher payload with very good imperceptibility will be a future scope in research on image steganography. As well as the survival of the proposed algorithm under other image processing operations like changes in brightness, cropping, resizing, etc., can also be considered for future research with improvements than existing algorithms results.

References

1. Bajpai S, Saxena K (2016) Evaluating data compression and image steganography. Int J Sci Eng Appl Sci 2(4):265–270
2. Yadav V, Sharma P (2015) A review paper on steganography. Int J Adv Eng Res Dev 2 (5):955–958
3. Zhang Y et al (2016) A framework of adaptive steganography resisting JPEG compression and detection. special issue In: Paper-security and communication networks-Wiley online library
4. Kumar V, Kumar D (2017) A modified DWT-based image steganography technique. Multimedia tools application cross mark. Springer
5. Kaur S et al (2016) An efficient adaptive data hiding scheme for image steganography. In: Proceedings of the international congress on information and communication technology springer science, pp 371–379
6. Zhang Y et al (2017) Dither modulation based adaptive steganography resisting jpeg compression and statistic detection. Springer-multimedia tools application cross mark
7. Aulakh N, Kaur Y (2015) Increasing image compression rate using (DWT + DCT) and steganography. Int J Emerg Res Manage Technol 4(5):253–260
8. Atawneh S, Putra S (2013) Hybrid and blind steganographic method for digital images based on DWT and chaotic map. J Commun 8(11):690–699

drksujatha23@gmail.com

9. Goti V, Shah N (2017) Survey: image steganography and its techniques. Int J Adv Res Eng Sci Technol 4(4):576–581
10. Mali S et al (2012) Robust and secured image-adaptive data hiding. Elsevier Digit Sig Process 22:314–323
11. Mishra M, Rout N (2017) Copyright protection of images by reversible, invisible, and robust digital watermarking using 3rd and 4th level discrete wavelet transform. Int J Eng Technol 9 (2):1084–1094
12. Sharma G, Kumar V (2017) A novel technique for reversible information hiding. Adv Comput Sci Technol Res India Publ 10(7):2069–2078
13. Mohammad R, Ali H (2016) Blind steganography in color images by double wavelet transform and improved Arnold transform. Indonesian J Electr Eng Comput Sci 3(3):586–600
14. Tamanna, Sethi A (2017) Analysis and refinement of steganography techniques. Int J Comput Appl 170(8):9–13
15. Dhaundiyal M, Nikumbh S (2014) Color image steganography based on wavelet transform. Int J Comput Appl 31–34
16. Furqan A, Kumar M (2015) Study and analysis of robust DWT-SVD domain based digital image watermarking technique using MATLAB. In: IEEE International conference on computational intelligence & communication technology, pp 638–644
17. Adwan Y et al (2015) An enhanced Steganographic model based on DWT combined with encryption and error correction techniques. Int J Adv Comput Sci Appl 6(12):49–55
18. Das S et al (2017) An improved DCT based image watermarking robust against JPEG compression and other attacks. Int J Image Graph Sig Proc 9:40–50
19. Kumar G, Shrivastava P (2016) A Survey of various image compression techniques for RGB images. Int J Eng Sci Comput 5(6):4905–4910
20. Hwang R et al (2004) A lossy compression tolerant data hiding method based on JPEG and VQ. J Int Technol 5(2):171–178
21. Rashid T, Dagar S (2016) Steganography techniques: a review. Int J Innovative Res Technol 2(12):387–391
22. Zhang Y et al (2016) Joint JPEG compression and detection resistant performance enhancement for adaptive steganography using feature regions selection. Multimedia tools application cross mark. Springer

drksujatha23@gmail.com

Improved Google Page Rank Algorithm



Abhishek Dixit, Vijay Singh Rathore and Anchal Sehgal

Abstract This paper is based on a Search Engine ranking algorithm. It proposes the technique for improving the page rank algorithm. The work focuses on the change in page rank algorithm, which helps in reducing the time complexity. We have calculated the normalized page rank by using the median value as it reduces the calculation work and time complexity. The comparison has been done between both the algorithms, i.e., old PR algorithm and the new proposed PR algorithm. This work also focuses on the research to increase the rank of the website. Various hybrid approaches are used to increase the rank of the website.

1 Introduction

This paper describes the algorithm and its application of Search Engine called Google and Optimization methods for a website and analyzes its effectiveness in the perspective of the search engine results [1]. It covers the various categories of techniques, keywords, old and new page rank calculation technique, algorithm, comparison, and its implementation [2]. Due to the swift growth of the Internet and the rapid increase in the number of websites available, it has become a tough task to search for best-existing sites for the searcher. According to a recent study, there are about 3 million new websites launched over the Internet per month [3]. Another study had proven that "the percentage of first visit to a website which comes from

A. Dixit (✉) · V. S. Rathore
Jaipur Engineering College & Research Centre, Jaipur, Rajasthan, India
e-mail: abhishekdixit.cse@jecrc.ac.in; abhishekdixit2606@gmail.com

V. S. Rathore
e-mail: vijaydiamond@gmail.com

A. Sehgal
Arya Institute of Engineering and Technology, Jaipur, India
e-mail: anchal.6455@gmail.com

© Springer Nature Singapore Pte Ltd. 2019

V. S. Rathore et al. (eds.), *Emerging Trends in Expert Applications and Security*,
Advances in Intelligent Systems and Computing 841,
https://doi.org/10.1007/978-981-13-2285-3_63

drksujatha23@gmail.com

web search is more than 80%" and the percentage share of those around 16% of Google searchers go beyond the second or further pages of the search results, and 65% of the users hardly ever click on paid, inorganic or sponsored results [4].

2 Page Rank Algorithm

2.1 Traditional Page Rank Algorithm

The formula for calculating the Page Rank, which was described by two famous people named as Sergey Brin and Lawrence Page and in some well-known publications which are given as [5]

$$\mathrm{PR(A)} = (1-\mathrm{d}) + \mathrm{d}[\mathrm{PR(T1)/C(T1)} + \mathrm{PR(T2)/C(T2)} + \mathrm{PR(T3)/C(T3)} + \cdots + \mathrm{PR(Tn)/C(Tn)}$$

where

- PR(A) is called page rank of A page.
- PR(T1) is the page rank of a page linked with page A.
- C(T1) is the total number of outgoing links from page T1.
- PR(Tn) is the page rank of nth page which is somehow linked to A page.
- C(Tn) is the total number of outgoing links from page Tn.
- d is called as damping factor whose value can be set anywhere between 0 and 1, usually in most of the researches value of d is considered as 0.85.
- (1 − d) is a bit of probability, some called it as math magic as it is considered as "sum of page ranks of all web pages will be one." It also explains the concept that if a page has no backlinks still it will get a small page rank, i.e., 0.15.

Calculation of old page rank algorithm is given by the following steps, which are as follows:

Step 1 Page ranks called PR of all the interrelated pages can be derived by the following formula:

$$\mathrm{PR(A)} = (1-\mathrm{d}) + \mathrm{d}[\mathrm{PR(T1)/C(T1)} + \mathrm{PR(T2)/C(T2)} + \mathrm{PR(T3)/C(T3)} + \cdots + \mathrm{PR(Tn)/C(Tn)}]$$

where d = 0.85.

Step 2 Repeat step 1 until the page rank values of two consecutive iterations match each other or about the same.

The above formula is based on the random surfer model. It is necessary to mention that many outgoing links to a single page or link to itself page will be ignored. The first step is to initialize the same Page Rank value for all the pages.

drksujatha23@gmail.com

The Page Rank transferring from a given page to all the targeted value or pages of its outgoing links in the next step is dispersed equally among whole outgoing links. It should be noted clearly that PR forms a probability distribution over all web pages, the sum of all the PR of all the web pages will come out only one [6].

Characteristics of Page Rank Algorithm

- It does not depend on the content of the website.
- It does not rank the whole website but page rank is calculated for each individual pages means whole website page rank is not calculated.
- It deals with the static web pages.
- Linking structure determine the page rank of all the pages.

2.2 Proposed Page Rank Algorithm

Step 1 Initially, it is assumed that page rank of all the web page to be 1

Step 2 Page ranks called PR of all the interrelated pages can be derived by following formula:

$$\mathrm{PR(A)} = (1-\mathrm{d}) + \mathrm{d}[\mathrm{PR(T1)/C(T1)} + \mathrm{PR(T2)/C(T2)} + \mathrm{PR(T3)/C(T3)} + \cdots + \mathrm{PR(Tn)/C(Tn)}]$$

Consider the value of d constant equal to 0.85.

Step 3 Calculate the median value (MV) of all the page rank values. Sort all the page rank values.
Let the number of values be N.
If N is odd,

$$\text{Median value (MV)} = \text{middle value.}$$

If N is even,

$$\text{Median value (MV)} = \text{mean of the two data values in the middle.}$$

Step 4 Normalize page rank of all the pages by the following formula:

$$\text{Normalized Page Rank (A)} = \text{Page Rank (A)/MV}$$

drksujatha23@gmail.com

Step 5 The assignment is done as

$$\text{Page Rank}\,(\text{A}) = \text{Normalized Page Rank}\,(\text{A})$$

Step 6 Process from Step 2 to Step 4 is repeated until PR values of two successive iterations are found the same.

3 Application Details

The newly developed page rank scheme is built upon normalization technique in that we have to choose a median value from different page rank values and then the normalized value is calculated by dividing the page rank value by median value and the process is repeated until the values of iteration matches the previous iteration values. Considering the following web graph.

As shown in Fig. 1, there are four web pages A, B, C, and D interconnected with each other. Page A has one backlink (incoming link), page B also has one backlink, page C three backlinks while page D does not have any backlink. We have applied the proposed page rank algorithm to this example and the results are compared with the results coming from traditional page rank algorithm.

We have proposed new page rank algorithm and comparison has been made in Table 3 about the number of iterations in the old PR algorithm in Table 1 and the number of iteration in the new proposed page rank algorithm in Table 2.

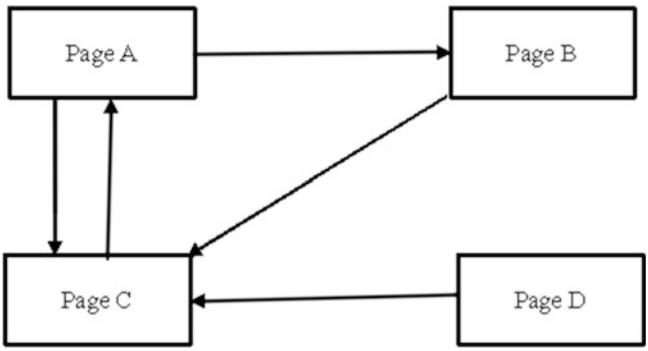


Fig. 1 Web graph showing the interconnection of different pages

drksujatha23@gmail.com

Table 1 Page rank values calculated from traditional page rank algorithm

Iteration	PR(A)	PR(B)	PR(C)	PR(D)
1	1	0.575	1.19	0.15
2	1.61	0.643	1.317	0.15
3	1.263	0.686	1.397	0.15
4	1.337	0.718	1.456	0.15
5	1.387	0.739	1.495	0.15
6	1.420	0.753	1.522	0.15
7	1.44	0.763	1.54	0.15
8	1.457	0.77	1.55	0.15
9	1.47	0.774	1.56	0.15
10	1.47	0.77	1.56	0.15

Table 2 Page rank values calculated from the proposed page rank algorithm

Iteration	PR(A)	PR(B)	PR(C)	PR(D)
1	1.27	0.730	1.51	0.19
2	1.30	0.69	1.42	0.17
3	1.30	0.70	1.43	0.166
4	1.30	0.70	1.43	0.16

Table 3 Comparison about the number of iterations

Number of iterations in traditional or old PR algorithm	Number of iterations in newly developed PR algorithm
10	4

4 Conclusion

In this newly developed scheme, the PR of all web pages are normalized by dividing all the page rank values by median value. It helps in decreasing the time complexity of the old PR algorithm. Computational research of page rank of all the pages signifies that new proposed scheme of calculating page rank is better than the traditional approach of calculating page rank in terms of time complexity. The result computed from the research shows that only four iterations are used by the proposed page rank algorithm as compared with traditional page rank algorithm which used 10 iterations for calculating the values. So we may conclude that the proposed scheme is a better approach for calculating page rank in terms of time complexity as well as computational savings.

drksujatha23@gmail.com

5 Future Scope

There is very large scope in future for the researchers who are planning to work on PR algorithm. Researchers can calculate the page rank of different pages based on some factor which defines the content value of the page and works on damping factor value can be done to improve the enactment of the newly developed algorithm.

References

1. Neshat HS (2011) Ranking of new sponsored online ads. IEEE
2. Su et al (2010) How to improve your Google ranking: myths and reality. IEEE
3. Search engine optimization. Retrieved from http://www.google.com/webmasters/docs/serach-engineoptimization-starter-guide.pdf
4. Ochoa ED (2012) An analysis of the application of selected search engine optimization techniques and their effectiveness on Google search ranking algorithms
5. Brin S, Page L (1995) The anatomy of a large-scale hyper textual web search engine. Stanford University, Stanford
6. Sanjay Kumar D (2015) A review paper on page ranking algorithms. Guru Jambheshwar University of Science and Technology (IJARCET)

drksujatha23@gmail.com

A Pioneering Encryption Technique for Images



C. Jeba Nega Cheltha and Rajan Kumar Jha

Abstract Security is the process that protects facts from accidental attacks. Now we are vastly depending on the Internet. We are conveying lot of facts via network. However, in numerous cases there is very less assurance for security. In many examples like military, industries, universities, medical fields, etc., conveying lot of images through communication channel. Protection is greatly required in countless fields in the present days. In our projected exertion, we are using CEILIDH method for encrypting and decrypting images. CEILIDH is a communal key cryptography scheme. It uses asymmetric input, in which both the dispatcher and beneficiary will be using secretive key and both of them know the communal key. For giving out communal key secretly, a honey encryption practice is used in our paper.

Keywords CEILIDH · Honey encryption · Communal key · Asymmetric

1 Introduction

In the contemporary world, we are habitually dependent on the Internet. We are transmitting lot of information through Internet. Sometimes we are sending much protected information. Examples like military, industries, universities, etc., have highly secure information. As Internet uses increases, hackers are also increased. Now it is the time for us to concentrate much on security. Many algorithms and concepts were introduced to secure information.

In our paper, we too focus on the security of images. As protection, for image is very much necessary in many fields. In our paper, we are using CEILIDH method for encrypting and decrypting images. As CEILIDH uses asymmetric key both the

C. Jeba Nega Cheltha (✉) · R. K. Jha
Jaipur Engineering College and Research Centre, Rajasthan, India
e-mail: jebanegacheltha.cse@jecrc.ac.in

R. K. Jha
e-mail: rajanjha.cse@jecrc.ac.in

© Springer Nature Singapore Pte Ltd. 2019

V. S. Rathore et al. (eds.), *Emerging Trends in Expert Applications and Security*,
Advances in Intelligent Systems and Computing 841,
https://doi.org/10.1007/978-981-13-2285-3_64

drksujatha23@gmail.com

dispatcher and recipient will use their confidential key to encrypt as well to decrypt. However, both the dispatcher and the recipient should know the communal key. To share communal key securely, honey encryption is used. If any hacker tries to hack the information using brute force attack, then the honey encryption will give false information. However, the hacker will assume the bogus information as original data. So to send communal key securely honey encryption technique is used.

2 Existing System

The Sanjay Kumar and Sandeep Srivastava used Simplified Data Encryption standard in their paper for image encryption [1]. Manjula and Ravikumar used DES algorithm in their projected work [2]. Anup and Suchithra used Triple Des algorithm for image encryption in their paper [3]. Silva-García et al. used Triple DES algorithm for encrypting the image [4]. Brindha et al. explained the use of the symmetric algorithm for image encryption [5]. All the above papers used symmetric key cryptography, in which, both the dispatcher and recipient uses the similar key. In our proposed work, we used the asymmetric key. So that both the dispatcher and recipient will be using a different key.

Brute force assault is an assault in which an assailant involves frequent decryption using casual keys [6]. If any aggressor tries to decrypt facts, honey encryption will send counterfeit fact for each erroneous presumption of the secret word. Therefore, that aggressor will presume that counterfeit facts as original fact. Even though both the dispatcher and recipient uses the dissimilar secretive key, they use the common communal key, which should be, well known to both the dispatcher and recipient. To share this communal key securely as well to avoid brute force attack, honey encryption technique is used in our proposed work.

3 Proposed Method

In our proposed work, two concepts are used. The image from the dispatcher will be encrypted by CEILIDH. The recipient will decrypt the message. For encryption and decryption, CEILIDH is used. As CEILIDH uses asymmetric key both the sender and receiver use private key to encrypt and decrypt image. However, both the dispatcher and recipient know the communal key. If they share the communal key, using Internet there is a possibility to hack the secret code. So to avoid this, honey encryption technique is used. If any aggressor tries to decrypt facts, honey encryption will send forged information for each flawed opinion of the clandestine word. Therefore, that provoker will assume that forged particulars as original fact. Both the algorithm are explained as follows.

drksujatha23@gmail.com

3.1 CEILIDH

It uses communal key based on the isolated logarithm problem. Alice Silverberg and Karl Rubin initiated this CEILIDH in 2003 [7]. The benefit of this CEILIDH is that it uses the small size of the key. CEILIDH algorithm explained below.

Let Pr be a prime power. Choose an integer int such that the torus Tor has a plain balanced parameterization, $\phi_{int}(q1)$ divisible by a big prime l1, where ϕ_{int} is cyclotomic polynomial.

Let m1 = ϕ(int) where ϕ is the Euler function. Let p: $\mathrm{Tor}_{int}\left(\mathrm{F1}_{q1}\right) \rightarrow \mathrm{F1}_{q1}^{m1}$ be a birational map. Choose $\alpha \varepsilon \mathrm{Tor}_{int}$ of order l1 and let g1 = p1(α).

Key agreement of CEILIDH is as follows. Let dispatcher chooses a random number a1 (mod $\phi_{int}(q1)$). Dispatcher computes $P1_A = p1(\Psi(g1)^{a1}) \varepsilon F1_{q1}^{m1}$ and sends it to the recipient and the recipient chooses a random number b1(mod $\phi_{int}(q1)$). Recipient computes $P1_B = p1(\Psi(g1)^{b1}) \varepsilon F1_{q1}^{m1}$ and sends it to the dispatcher. Now the dispatcher computes $p1(\Psi(P1_B)^{a1}) \varepsilon F1_{q1}^{m1}$ and the recipient computes $p1(\Psi(P1_A)^{b1}) \varepsilon F1_{q1}^{m1}$.

In encryption, the message M1 is an element of $\varepsilon F1_{q1}^{m1}$. The dispatcher chooses a random integer k in the range $1 < k1 < l1-1$. Also, the dispatcher computes $\curlyvee = p1(\Psi(g1)^{b1}) \varepsilon F1_{q1}^{m1}$ and $\ni = p1(\Psi(M1)\ \Psi((P1_A)^{k1})) \varepsilon F1_{q1}^{m1}$. The dispatcher sends the ciphertext ($\curlyvee$, $\ni$) to the recipient. Now, the recipient computes the message M1 as $p1(\Psi(\ni)\ \Psi((\curlyvee)^{-a)})$.

3.2 Honey Encryption

Safety starts with authenticating; commonly we are using username and code word. Most of the code word used in the contemporary globe is not brawny. Numerous users are using a weak code word.

Honey Encryption (HE) is a category of facts encryption, which produces code text [8]. Ari Juels and Thomas Ristenpart introduced honey encryption in 2014 [9, 10]. In honey encryption, if any aggressor tries to decrypt with an erroneous code word it will show a believable look [2]. In this proposed work, honey encryption is used to send a communal code which CEILIDH.

A honey encryption method HE1 = ($H1_{Enc}$, $H1_{Dec}$) is an algorithm. $H1_{Enc}$ indicates encryption by honey encryption and $H1_{Dec}$ indicates decryption by honey encryption. Assume communication space as M1, seed space K1 which is the space of every n1 bit binary series for a few agreed n1.

Message M1 encloses every feasible message. Encryption $H1_{Enc}$ obtains input key K1 $\in$ K1, message M1 $\in$ M1, a few standardized chaotic bits, and yields a code text C1. X1 indicates key and R1 indicates standardized arbitrary bit. Chart message M1 to seed space K1 from initial to end by utilizing allocation-transforming encoder.

drksujatha23@gmail.com

Honey encryption is not easy to pertain since the space of plaintext is extremely huge. The quantity of the seed sequence of m1 is directly relative to believable m1 in the communication space M1.

Allocation transforming code demands the cumulative distribution purpose of M1 and in succession of sorting messages. By this data, we know how to hesitate on the cumulative probability sequence corresponding to the message m1 and design it to the matching percentile seed range in K1.

Honey encryption algorithm is explained as follows:

$H1_{Enc} \leftarrow$ Enc (X1, M1),
K1 $\leftarrow$ \$ encode (M1),
R1 $\leftarrow$ \$ {0, 1} n1,
K1″ $\leftarrow$ H1 (R1, X1),
C1 $\leftarrow$ K1″ $\oplus$ K1

where \$ indicates that $H1_{Enc}$ may perhaps exploit different amount of standardized chaotic bits.

Honey decryption algorithm $H1_{Dec}$ confines input as a key K1 $\in$ K1, and output as a message M1 $\in$ M1.

Decryption of honey encryption is as follows:

$H1_{Dec} \leftarrow$ Dec (X1, (R1, C1)),
K1″$\leftarrow$ H1 (R1, X1),
K1 $\leftarrow$ C1 $\oplus$ K1″,
M1 $\leftarrow$ decode (K1),
Return M1



Fig. 1 Original image

drksujatha23@gmail.com

In this anticipated work, we can choose any image. The preferred image will be encrypted using the projected method. Figure 1 shows the original image before encryption, which is the image captured by us [11] for testing our proposed work, Fig. 2 shows the encrypted image. To decrypt the image we used the projected method and Fig. 3 shows the decrypted image using CEILIDH [11].

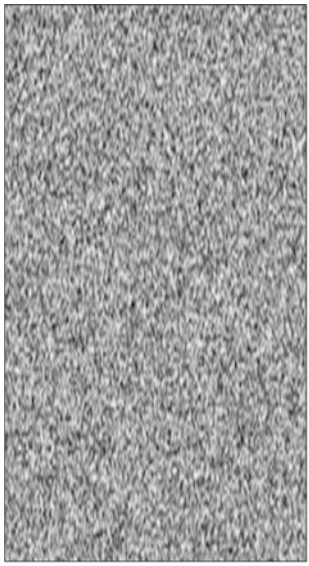


Fig. 2 Encrypted image



Fig. 3 After decryption

drksujatha23@gmail.com

4 Conclusion

We have proposed a novel method for image encryption and decryption using CEILDIH in this paper. CEILIDH is a communal key cryptography scheme. It uses asymmetric input, in which both the dispatcher and beneficiary will be using the secretive key and both of them know the communal key. For giving out communal key secretly, honey encryption practices are proposed in our paper. Honey encryption avoids brute force attack also, it fools hacker.

Acknowledgements I would like to express gratitude to Almighty and all those who supported to complete this projected work. In addition, I would like to mention that in Figs. 1 and 3 there is a human image, which was captured by us [11] to test our anticipated work. While decrypting the image using our proposed work, we successfully obtained Fig. 3. Especially, I thank the child who posed for this image to complete our proposed work successfully.

References

1. Kumar S, Srivastava S (2014) Image encryption using simplified data encryption standard (S-DES). Int J Comput Appl 104(2):(0975–8887)
2. Manjula KG, Ravikumar MN (2016) Color image encryption and decryption using DES algorithm. Int Res J Eng Technol (IRJET) 03(07). e-ISSN: 2395-0056. www.irjet.net p-ISSN: 2395-0072
3. Anup R, Suchithra R (2017) Image encryption using triple DES algorithm. Imperial J Interdisc Res (IJIR) 3(5). ISSN: 2454-1362
4. Silva-García VM, Flores-Carapia R, López-Yañez I, Rentería-Márquez C (2012) Image encryption based on the modified triple-DES cryptosystem. Int Math Forum 7(59):2929–2942
5. Brindha K, Sharma R, Sain S. Use of symmetric algorithm for image encryption. Int J Innovative Res Comput Commun Eng
6. Adleman LM, Rothemund PWK, Roweis S, Winfree E (1996) On applying molecular computation to the data encryption standard. In: Proceedings of the second annual meeting on DNA based computers, Princeton University, 10–12 June 1996
7. Rubin K, Silverberg A (2003) Torus-based cryptography. CRYPTO 349–365
8. Simonite T (2014) Honey encryption will bamboozle attackers with fake secrets. MIT technology review. Accessed 30 Jan 2014
9. Vinayak PP, Nahala MA (2015) Avoiding brute force attack in MANET using honey encryption. IJSR 4(3)
10. Juels A, Ristenpart T (2014) Honey encryption: encryption beyond the brute-force barrier. IEEE Secur Priv 12(4):59–62
11. https://www.dropbox.com/home/Camera%20Uploads?preview=2016-01-17+13.13.37.jpg

drksujatha23@gmail.com

A Pedestrian Collision Prevention Method Through P2V Communication



JinHyuck Park, ChoonSung Nam, JangYeol Lee and DongRyeol Shin

Abstract Today, research and commercialization about the autonomous vehicle are being progressed. Among them, the most popular issue is self-driving. To support self-driving, a vehicle has to know the pedestrian's location. Pedestrians have a smartphone with BLE communication. Thus, this paper proposes Bluetooth Lower Energy (BLE) communication-based service that recognizes the pedestrian and sends a warning message to the vehicle. It can be easily usable because it is easy to find a device that uses BLE communication such as a smartphone.

1 Introduction

While the research and commercialization for autonomous vehicles are in progress, the largest issue has been the research for safe driving. Autonomous vehicles must be able to sense the danger and be able to control the vehicle itself according to a pedestrian's location or expected motion. Therefore, it should receive external information from pedestrians [1].

The communication method needs network technology between vehicles and other devices, and Vehicular Ad hoc Networks (VANETs) are required to make it possible. VANETs can broadly be divided into two communication methods. First, through Vehicle to Vehicle (V2V) communication, one vehicle is able to predict

J. Park
Department of Cyber Security, Kyungil University, Gyeongsan, South Korea
e-mail: vkqxkr@gmail.com

C. Nam · J. Lee · D. Shin (✉)
Department of Electric & Electrical Computer Engineering,
Sungkyunkwan University, Suwon, South Korea
e-mail: drshin@skku.edu

C. Nam
e-mail: namgun99@gmail.com

J. Lee
e-mail: ljygo2005@skku.edu

© Springer Nature Singapore Pte Ltd. 2019

V. S. Rathore et al. (eds.), *Emerging Trends in Expert Applications and Security*,
Advances in Intelligent Systems and Computing 841,
https://doi.org/10.1007/978-981-13-2285-3_65

drksujatha23@gmail.com

other vehicles' movement so that it may recognize the danger in advance [2]. The second technology which is Vehicle to Infrastructure (V2I) is able to provide road conditions through information exchange between infrastructures such as vehicles or Road Side Unit (RSU) [3, 4]. However, the communication method of V2V and V2I of VANETs mainly consists of receiving safe messages as intersection movement assist and left turn assist between cars implying that there is no way to trace the position data of the pedestrian. Thus, to obtain the pedestrian's data, it is required to have the technology that is capable of both transmission and reception between a vehicle and a pedestrian.

Bluetooth Low Energy (BLE) communication can be used to provide the pedestrian's position data [5]. Basically, if a pedestrian possesses a smart device, the device will produce BLE and be used as the communication between the pedestrian and the vehicle based on the pedestrian's data. In addition, since BLE does not require a connection method through pairing, it is possible to transmit data through the broadcast. Yet, if the vehicle is operated carelessly based on random pedestrian's data, a traffic collision will occur by unnecessary vehicle operation. In other words, the communication between the pedestrian and the vehicle needs a method that can extract the pedestrian's data in the situation of vehicle operation.

Therefore, this paper proposes a method to set the safe zone according to the speed and position of the vehicle by using the BLE communication between the pedestrian and the vehicle to ensure both the danger detection and safe driving.

The remaining paper is organized as follows: Sect. 2 explains the current research in location recognition or service with related work. Section 3 suggests the communication method and algorithm between the pedestrian and the vehicle. Finally, Sect. 4 shows conclusion and future research directions.

2 Related Work

2.1 Communication Method Between the Pedestrian and the Vehicle

VANETs for the Communication Between Vehicles and Infrastructures

VANETs are actively being used in various application researches and commercial field starting from communication model. Among them, commercialization of V2V is proceeding mostly. V2V is defined as the communication between vehicles and aims to increase the safety between them and help drivers to drive safely. According to the report, "Vehicle-to-Vehicle communication: Readiness of V2V Technology for Application" [6] published by National Highway Traffic Safety Administration, V2V is capable of increasing both road and drivers' safety. Furthermore, the report defines some applications through scenarios, and it explains the importance and the necessity of V2V. However, the safe scenarios through V2V communication consist of merely the communication between vehicles, not environments, and more

drksujatha23@gmail.com

importantly pedestrians. Moreover, although it defines application models such as P2V and Pedestrian in Roadway Warning (PRW), there is a weakness that it does not propose the definition of accurate communication and the technology that solves a danger between the pedestrian and the vehicle and other safety issues. Therefore, the communication method between the pedestrian and the vehicle is required.

iBeacon for P2V Communication

iBeacon [7] is a Near-Field Communication (NFC)-based service device introduced by Apple Inc. It is the service device using NFC between users and adopting the technology named Bluetooth Low Energy (BLE). BLE is a communication technology which consists of low packet parameter such as low transmitted power and received power to reduce energy consumption. The existing Bluetooth communication technology communicates NFC through the connection between Bluetooth enabled devices called pairing. iBeacon, on the other hand, is designed to allow devices capable of BLE communication to communicate rapidly using broadcast-type transmission without any kind of connection. BLE communication technology is the method that is able to provide vehicles with pedestrian's information without any connection. The method to measure distance with iBeacon can estimate the distance between the smart device and iBeacon by calculating Received Signal Strength Indication (RSSI) value and Tx power value. The communication method between iBeacon and the smart device is shown in Fig. 1.

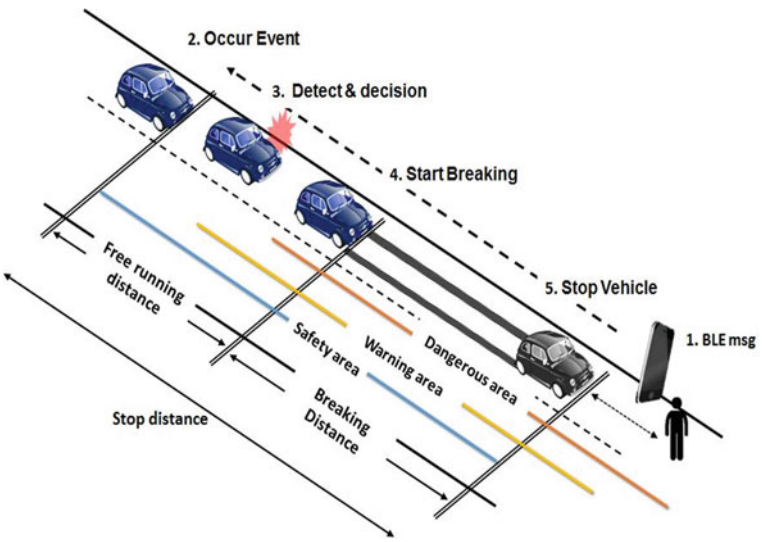


Fig. 1 Safety area architecture through BLE

drksujatha23@gmail.com

When the smart device enters the transmission area of the iBeacon, the smart device can receive the iBeacon data and check the information of the device. Therefore, this communication is possible if a Bluetooth device is installed in a vehicle.

3 Danger Recognition Method

To provide a hazard signal between the pedestrian and the vehicle, they must be able to perceive their relative data with respect to their current position, speed, and direction. In order to do it, they should be able to exchange communication in the form of P2V as shown in Fig. 2. The vehicle finds the point where it can brake by itself by calculating its own speed, the distance between pedestrians based on the pedestrian's location. This point includes a safe area where the vehicle can maintain its existing driving without special actions, a warning area where speed reduction and a change of direction must be made within certain time, and a danger area where the vehicle control itself immediately without the driver's decision. Depending on the area, the vehicle generates events such as driving, warning, braking, etc. The pedestrian is able to detect a situation in advance in which the pedestrian is located on the roadway based on the data of the vehicle received, and deal with it. Therefore, both the pedestrian and the vehicle prevent traffic collision by providing danger recognition.

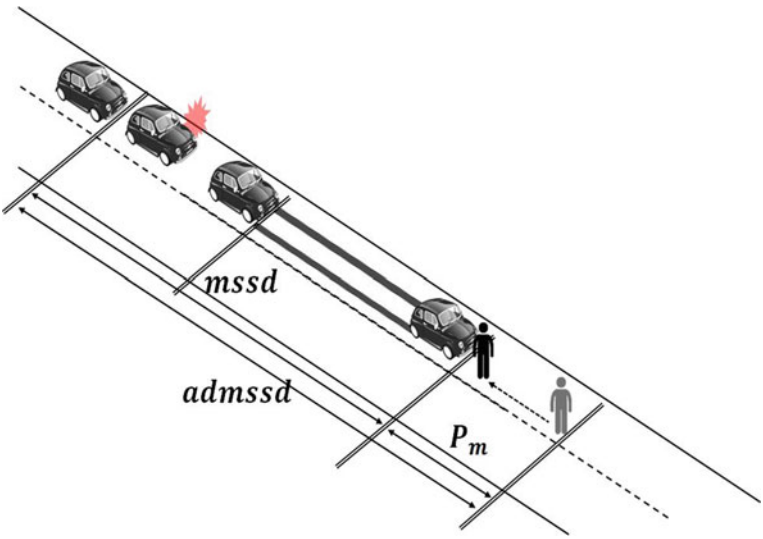


Fig. 2 Braking distance: admssd for preventing collision

drksujatha23@gmail.com

3.1 Setting up Safety Area

The safe driving zone is defined as the area outside the distance until the vehicle in motion comes to a complete stop at its current speed. When the vehicle receives a message from the pedestrian through the smart device, a total of the distance traveled by the vehicle until it receives the message, until the driver steps on the brake from receiving the message, and for speed reduction from applying the brake should be less than the distance between the pedestrian and the vehicle.

In order to obtain the stopping distance, the definition of the previously defined stopping distance of the vehicle should be considered. The description of the stopping distance is shown in Fig. 2. The stopping distance can be divided into two categories: free running distance, braking distance. Free running distance means the amount of time either the driver or the vehicle decides to stop and starts to apply the brake when the vehicle detects a dangerous object or the pedestrian to appear on the roadway. Breaking distance is the distance from the point where the vehicle starts to work the brake to stopping completely. The sum of these two distances is the vehicle's stopping distance.

Equation (1) is the expression for calculating the basic breaking distance. *mssd* stands for Minimum Stop Sight Distance. This allows the total stopping distance of the vehicle to be calculated. In Eq. (1), V is the design speed of the road (or vehicle speed) and t is the driver's recognition response time (or computation time) and f is the coefficient for the road friction.

$$\text{mssd} = \frac{tV}{3.6} + \frac{V^2}{256f} \tag{1}$$

Equation (2) is the formula for the distance that the vehicle travels relative to the time when the pedestrian enters the road. Because the formula above shows the total stopping distance, the hazardous area can be established.

$$\text{Pm} = \left(\frac{\frac{mssd}{V}}{3600} \times 1.8\right) + 1 \tag{2}$$

$$\text{admssd} = \text{mssd} + \text{Pm} \tag{3}$$

3.2 Results Analysis

Figure 3 shows minimum safety distance between a pedestrian and a vehicle by vehicle's speed, when the driver's cognitive response time is set to a maximum of 1.5 s. If the speed of the vehicle is more than 100 km, the driver has to know

drksujatha23@gmail.com

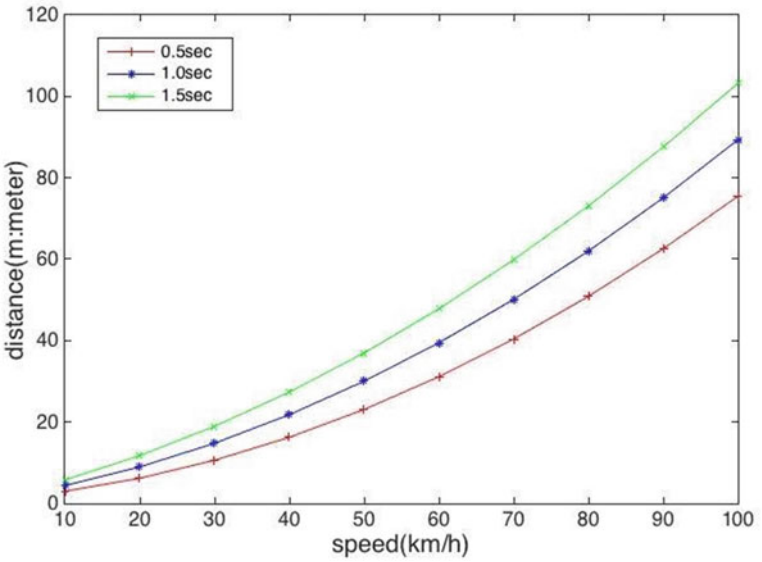


Fig. 3 Safety distance by cognitive response time of accident alarm

accident alarm before 100 m from a pedestrian. However, if the cognitive response time of the driver is shortened to 0.5 s, the safety distance is about 70 m. Thus, it is possible to prevent a vehicle accident by constructing the proposed safe area system using BLE communication.

4 Conclusion and Future Research Plan

This paper presented a method to prevent accidents of vehicles and pedestrians using BLE communication. To achieve this, the distance is measured by exchanging the location information of the vehicle and the pedestrian, and the drivers know the alarm about the stopping distance.

Future research will measure the possible distance for accident prevention through simulation.

Acknowledgements This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korean government (MSIT) (No. R0113-15-0002, Automotive ICT-based e-Call standardization and after-market device development).

drksujatha23@gmail.com

References

1. An KH (2013) Technology trends of self-driving vehicles. Electron Telecommun Trends (ETRI) 35–44
2. Lin YW (2010) Routing protocols in vehicular ad hoc networks: a survey and future perspectives. J Inf Sci Eng (IIS) 26:913–932
3. Qian Y (2008) Design of secure and application-oriented. In: Vehicular technology conference 2008. IEEE, pp 2794–2799
4. Karagiannis G (2011) Vehicular networking: a survey and tutorial on requirements, architectures, challenges, standards and solutions. Commun Surv Tutorials. IEEE, 584–616
5. Frank R (2014) Bluetooth low energy: an alternative technology for VANET applications. In: Wireless on-demand network systems and services 11th annual conference. IEEE, pp 104–108
6. Harding J (2014) Vehicle-to-vehicle communications: readiness of V2V technology for applicaiton. U.S. Department of Trasportation, National Highway Traffic Safety Administration
7. Cavallini A (2014) iBeacons Bible 2.0. http://meetingofideas.wordpress.com/

drksujatha23@gmail.com

Summarization Using Corpus Training and Machine Learning



Vikas Kumar, Tanupriya Choudhury, A. Sai Sabitha and Shweta Mishra

Abstract Automatic summarization could be used for finding useful data from a given speech or text. Automatic summarization requires a machine learning approach to find the most suitable sentences to be included in the summary. Since summarization is a human process, it requires a human-like thinking approach from a machine. Summarization could be used for automatically finding out the main highlights of a given article or speech. First, we start with sentence extraction. Then, we use the corpus to find relevant patterns or features according to which we could rank a sentence. We train a Naive Bayes Classifier according to those features. Then, we perform tests on the Naive Bayes Classifier for finding scores of each sentence. The summary from the original text is produced using a certain compression rate according to which the machine selects the n-best sentences. A neural network model was also trained to compare results. The paper ends with an evaluation of the procedure's accuracy by testing it against different test cases of various lengths and varying compression rates.

Keywords Text mining · Naive Bayes classifier · Automatic summarization Corpus mining · Artificial neural · Network

V. Kumar · T. Choudhury (✉) · A. Sai Sabitha · S. Mishra
Amity University Uttar Pradesh, Noida, Uttar Pradesh, India
e-mail: tchoudhury@amity.edu; tanupriya1986@gmail.com

V. Kumar
e-mail: vkumar111223@gmail.com

A. Sai Sabitha
e-mail: saisabitha@gmail.com

S. Mishra
e-mail: backtoshweta@gmail.com

© Springer Nature Singapore Pte Ltd. 2019

V. S. Rathore et al. (eds.), *Emerging Trends in Expert Applications and Security*,
Advances in Intelligent Systems and Computing 841,
https://doi.org/10.1007/978-981-13-2285-3_66

drksujatha23@gmail.com

1 Introduction

Automatic summarization has several applications. It can be used in offices while working, office meetings, government centers, call centers, presentations, news analysis. The technology is only preferable when the amount of data to be processed is high. Even though the technology can be extremely helpful, people always prefer to read or listen to the whole text to have a deeper understanding. But, if we need a fast review of a large document and if the number of documents is too high, we use text mining to find only the relevant data.

Speech summarization needs a different approach from text summarization. A piece of text is usually well written with a special care for grammatical mistakes, repetition, and redundancy. But, the above errors occur too often in speeches along with some others such as nervousness, stammering, and mispronunciation. Factors for judging the efficiency of the algorithm includes the capacity of the algorithm to reduce the above-mentioned errors. Two steps have been followed in this paper for speech summarization, i.e., feature selection and training using Naïve Bayes and ANN.

2 Related Work

Recent research in this area takes advantage of various sentence features such as word frequency, positioning, thematic words, length, TF-IDF, and so on [1]. By using a machine learning approach, it has become possible to learn the properties of a text or rules of sentence formation from a corpus of documents.

There are two main approaches followed for sentence summarization- supervised and unsupervised techniques. Supervised techniques rely on the previous speech-summary pairs, whereas unsupervised techniques generate a summary based on the properties or features of the given text.

The amount of resources and time spent in Automatic Speech Summarization is huge. In fact, much research has been done in this field. But still, we have not been able to perfect the art of summarization due to human errors while delivering a speech. Though we have been able to achieve a good result with around 50–60% accuracy for any given speech [1]. Summarization using machine learning (Using corpus method) has proved to be more effective. The accuracy of such machines start with 60–70% accuracy and achieve up to 94% accuracy [2].

Neural networks are a new field of research. Some studies have also been conducted for sentence selection according to weightage using neural networks with a really high accuracy. (Ron Brandow, 772–879) Since summarization is highly dependent on sentence selection using several factors which may or may not be interdependent, it can be implemented using neural networks.

Kaikhah [3] has explained different procedures for corpus analysis along with different training procedures such as Naive Bayes, C4.5 Trainer, and Microsoft Speech Summarizer. He obtained a maximum of 50% accuracy with NB classifier.

drksujatha23@gmail.com

Table 1 Factors for deciding the sentence score

Feature f1	N-gram percentage of the sentence
Feature f2	% of words relating to the given keywords
Feature f3	% of content words (nouns, adjectives, verbs)
Feature f4	Positional score of the sentence in the document
Feature f5	Length of the sentence
Feature f6	Sentence to sentence cohesion

3 Feature Analysis

Let T be a text of Si sentences where i = 1, 2, 5, 6… till N. So, the sentence score (SSc) for a sentence can be calculated using factors stated in Table 1.

These factors are the features used in Artificial Neural Network. Using supervised training, we will try to determine the dependency of each factor for determining the sentence score of the sentence.

3.1 Procedure for Corpus Analysis

Corpus analysis is the analysis of patterns form a very large text and mining for a specific pattern from that text. It is a concept of machine learning. An in-depth

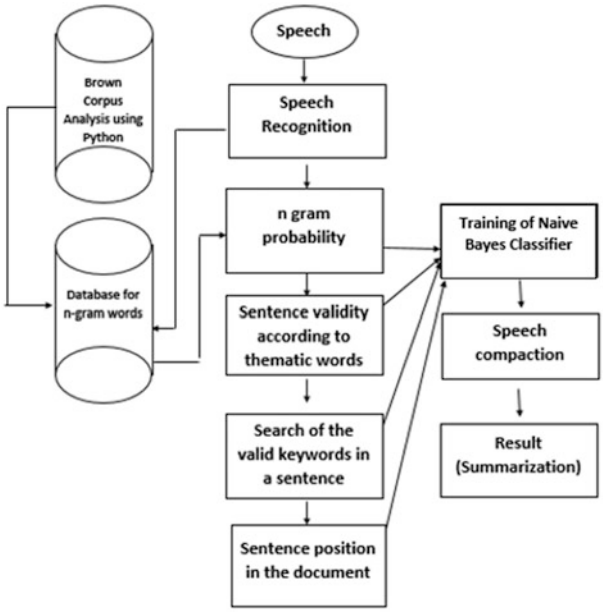


Fig. 1 Procedure for sentence summarization using corpus analysis

drksujatha23@gmail.com

analysis of linguistic patterns can yield some great results such as n-gram words (unigram, bigram, and trigram words). It can also tell us about the most frequent words used in newspapers, research papers, and other forms of written and electronic media. Such information can be used to predict the sentence, check the grammar accuracy, and find the relevancy of data according to the current topic. The procedure for Corpus Analysis is shown in Fig. 1.

3.2 Brown Corpus

The Brown University Standard Corpus is a library formed using Modern English words and is helpful to find relevant data. The Brown Corpus is freely available in the nltk module of Python. Brown Corpus has categories of text such as news, editorial, adventure, fiction, government, etc. The Brown Corpus is used in this paper to build a library of n-gram words (discussed below).

3.3 N-Gram Probability (Factor F1)

N-gram words in a text can be found with ease. For e.g., in the text 'A quick fox jumps over a lazy dog'

Unigrams—[A, quick, brown, fox…]

Bigrams—[(A, quick), (quick, brown), …]

Trigrams—[(A, quick, brown), (quick, brown, fox)…]

The n-gram combinations for a text are shown above and can be found directly from the given text. But for finding the n-gram probability of a given text, we need to find the n-gram patterns from a large dataset. In this paper, we observe the trigram possibilities and the number of times a specific possibility occurs using the brown corpus. Using this way, we train the database using a huge collection of text. We find the possible unigrams and trigrams [4].

3.4 Relating Keywords (Feature F2)

Relating keywords is a useful feature if the user is looking for a summary according to his desire. In that case, the user can define a few keywords. The machine should be trained such that, if the user chooses to define a few keywords for summarizing, the machine should give priority to statements in the summary relating to those keywords.

drksujatha23@gmail.com

3.5 Sentence Cohesion Score (Feature F6)

This feature determines how a sentence is related to other sentences. It does not only mean lexical analysis but also semantic analysis. For e.g., make and create are synonyms and if they appear in two different sentences, they make the sentences similar to each other.

For finding sentence cohesion score, we use the wordnet corpus from the python nltk module. The corpus allows us to compare two words and find the similarity score ranging from [0, 1]. The similarity score is based on spellings as well as meanings of the words. Using this, we perform a word to word comparison between two sentences and find the max similarity score for two sentences. We use this algorithm for every pair of sentences and then average the similarity scores obtained by comparing a sentence to every other sentence (Ron Brandow, 772–879).

(1) Target sentence is represented in a word list S
(2) Every other sentence is stored as a word list Si in a list SEN where SEN = [S1, S2, S3…, Si]
(3) Use NLTK.POS_TAG on every word in list S and all lists in SEN and remove verbs, conjunctions, articles
(4) Use the Wordnet module from python.corpus
(5) SIM_SCORE = 0
(6) For each Si in SEN:

    For each WORD1 in S:

    For each WORD2 in Si:

    Store WORD1.similarity(WORD2) in SIM_LIST

(7) M = MAX(SIM_LIST)
(8) SIM_SCORE = SIM_SCORE + M

$$f6(S) = \text{SIM\_SCORE}/length(SEN) \quad (1)$$

4 Intermediate Results

After Corpus Training, we evaluate the scores of sentences according to the six discussed features. These scores are fed into the Naive Bayes Classifier as well as ANN classifier to train it as discussed later. The scores of 10 sample sentences from an article of Hindustan Times are shown in Table 2. During the training, the classifier will try to determine the dependency of every feature on the consolidated sentence score. Once the machine is trained, we will use the machine on test data to get the final results, i.e., the summary of the text.

drksujatha23@gmail.com

Table 2 Sentence scoring according to features

Sentence number	f1 N-gram probability	f2 Keywords found	f3 Content words	f4 Positional score	f5 Relative sentence length	f6 Sentence cohesion score
1	0.2231	0.083	0.233	0.833	0.344	0.495
2	0.0588	0.045	0.186	0	0.427	0.804
3	0.1856	0	0.583	0.833	0.170	0.612
4	0.0403	0.08	0.330	1	0.515	0.77
5	0.0474	0	0.5	0.333	0.615	0.873
6	0.0677	0.031	0.318	0.333	0.816	0.761
7	0.0938	0	0.25	1	0.305	0.596
8	0.2368	0	0.375	0.25	0.336	0.596
9	0.0862	0.038	0.292	0.25	0.650	0.775
10	0.1154	0.023	0.276	0.166	1	0.787

4.1 Correlation Between Features (Spearman Rank Test)

Extraneous features mean those features which are linearly or nonlinearly dependent on other features. Removing such features is important because:

1. It increases ambiguity.
2. It increases runtime of machine.
3. The most important reason is that for a Naive Bayes Classifier to work properly, the features should be independent of each other. If some of the features are correlated, the functionality of the classifier is compromised [5].

For performing Spearman Rank Test, a set of 137 sentences and 38 paragraphs from 4 different Wikipedia articles were tested. The coefficients between two datasets are shown in Table 3. There is no strong correlation between any of the two features. Hence, we can use all six scores in the classification stage (Fig. 2).

Table 3 Coefficient of Spearman rank test (high correlation for sentence cohesion)

	f1	f2	f3	f4	f5	f6
f1	–	−0.04	0.35	0.054	−0.10	−0.12
f2	–	–	0.23	0.008	−0.08	0.361
f3	–	–	–	0.131	−0.11	0.178
f4	–	–	–	–	−0.15	−0.14
f5	–	–	–	–	–	0.477

drksujatha23@gmail.com

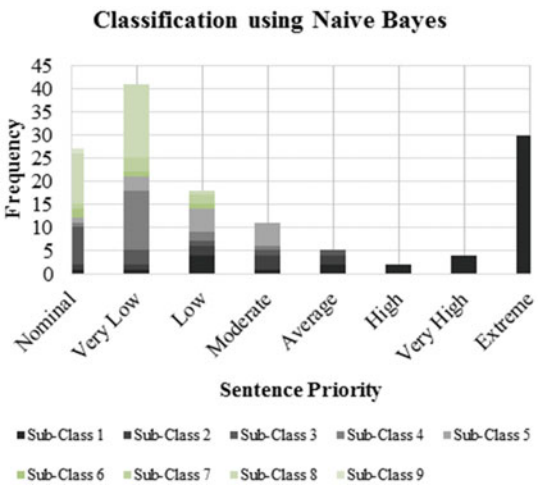


Fig. 2 Sentence scoring according to Naive Bayes (from test data)

5 Artificial Neural Network

We use yet another procedure for classification which is neural networks which can classify data items with nonlinear dependency. Unlike earlier case, we classify the sentences in just two categories (sentences to be included and sentences not to be included) in the summarization. For testing, we use 0 for sentences to be included and 1 for sentences not to be included. The neural network used in this case is:

1. **First layer (Input layer)**—6 nodes representing features f1, f2, f3, f4, f5, f6
2. **Second layer (Hidden layer)**—3 nodes with ReLU (Rectified Linear Unit) activation

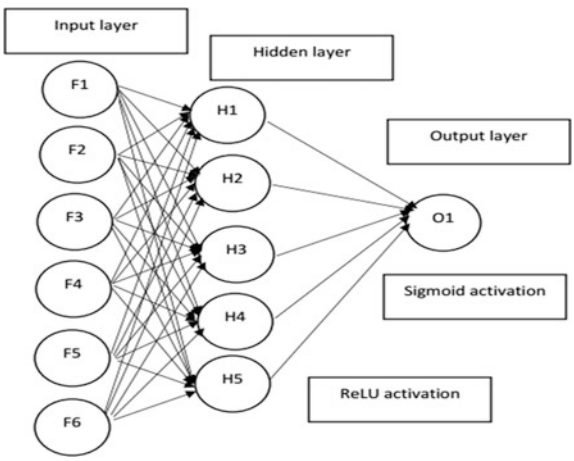


Fig. 3 Neural network architecture

drksujatha23@gmail.com

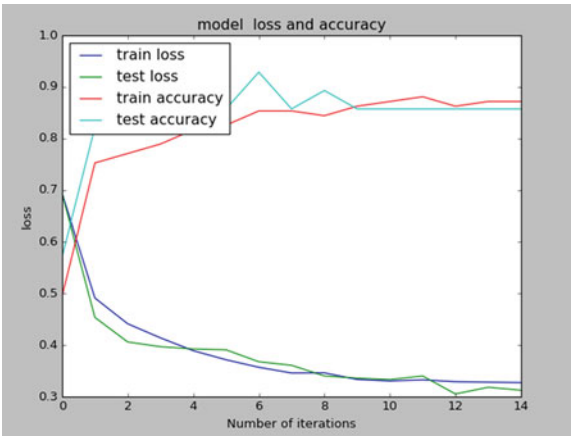


Fig. 4 Accuracy of the experiment

Table 4 Efficiency analysis of Naive Bayes and ANN

Trainer	10% CR (%)	20% CR (%)	30% CR (%)	50% CR (%)
Train (NB)	89	88	88	88
Test (NB)	80	80	80	80
Train (ANN)	93	95	88	86
Test (ANN)	87	85	85	83

3. **Third Layer (Output layer)**—1 node with sigmoid activation and output being either 0 or 1. If output is 0, then sentence should be included. If output is 1, then sentence should not be included (Fig. 3)
4. **Loss Function**—Binary Cross-Entropy is used to find the difference between the actual output and the target output and the weights are corrected according to that. The loss should decrease over a number of iterations for the neural network to be efficient as shown in Fig. 4. Efficiency Test.

The efficiency of both Naive Bayes, as well as ANN classifier, was found out using speech compression rate. Several tests were performed on different data sizes. Compression rates let the machine know how much data has to be removed from the original speech. The accuracy of the machines can be seen in Table 4. CR stands for compression rate. NB stands for Naive Bayes and ANN stands for Artificial Neural Network.

drksujatha23@gmail.com

6 Conclusion

Both the training methods performed well in the text summarization process. The Naive Bayes particularly proved to be efficient because it could successfully divide the sentences into 38 classes with increasing priority. Such high classification with such a high accuracy is beyond expectations. Using a neural network was also necessary since summarization is a process which is closely related to how a human thinks. Both the methods were almost equally efficient. The efficiency of Naive Bayes Theorem was around 75–85% for all test results and all compression rates while the accuracy of Neural Networks was around 83–90% for all test results. Though the trained neural network had higher fluctuation as compared to Naive Bayes Theorem. The efficiency produced is better than many previous summarization processes [5].

References

1. Joel Larocco Neto AA (2014) Automatic speech summarization using a machine learning approach. In: Symposium on artificial intelligence, pp 1–8
2. Padma Priya G, Duraiswamy K (2014) An approach for text summarization using deep learning algorithm. J Comput Sci 1:1–9
3. Kaikhah K (2013) Text summarization using neural networks. IBM J Res Devel, 2–7
4. H L (2002) The automatic creation of literature surveys. J Res Deve BM, 150–160
5. I M (1998) History of text summarization processes. MIT Press, pp 225–227

drksujatha23@gmail.com

Exploring Open Source for Machine Learning Problem on Diabetic Retinopathy



Archana Kumari, Tanupriya Choudhury and P. Chitra Rajagopal

Abstract Open-source operating system, as well as its packages, is more powerful and secure than the proprietary sources. In the proprietary source, software source code is not easily available because it is secret; by contrast in the open-source operating system source code is easily available, so any programmer can change the code and implement their ideas and modify it because of its openness. Also, one major advantage is that we do not need to spend a huge amount of money on the software. So, in this paper, we used open-source software for coding purposes and looked at the data available on the UCI machine learning repository on the diabetic retinopathy.

Keywords Open-source · Boosted decision tree · Neural network
Diabetic retinopathy

1 Introduction

Diabetes is one of the major concerns in the whole world and as per the WHO India will become the diabetic center of the world by 2030 [1]. Diabetes is known as silent killer of the person because of this many symptoms occurred in the whole body if it is not taken seriously. Complications occurred in a diabetic patient can be broadly classified into two categories [2]. They are:

A. Kumari (✉) · P. Chitra Rajagopal
Amity University Uttar Pradesh, Noida, Uttar Pradesh, India
e-mail: archanabtech1501@gmail.com

P. Chitra Rajagopal
e-mail: chitrargpnirmal@gmail.com

T. Choudhury
UPES, Dehradun, India
e-mail: tanupriya1986@gmail.com

© Springer Nature Singapore Pte Ltd. 2019

V. S. Rathore et al. (eds.), *Emerging Trends in Expert Applications and Security*,
Advances in Intelligent Systems and Computing 841,
https://doi.org/10.1007/978-981-13-2285-3_67

drksujatha23@gmail.com

1. Microvascular Complications: This creates issues related to the small blood vessel like problem with eyes (Retinopathy), nerves (neuropathy), and kidneys (nephropathy).
2. Macrovascular complications: This affects issues related to large blood vessel. Thus, it creates issues with heart and brain.

In this paper, detection of Diabetic Retinopathy (DR) [3] is shown using a machine learning technique. It generally appears, when diabetes mellitus persists for many years along with a certain group of the specific lesion in the retina. Lesion in retina tries to damage the eyes. If it is identified in the early stage then patients are rescued from vision loss. DR is one of the major causes of eyes loss in middle age people in developed and developing countries. According to WHO, in India, about 31.7 million people are affected by diabetes and it will be increased to 80 million by 2030. More than 80% of people who have diabetes longer than 15 years develop the DR that finally leads to the loss of vision [4, 5].

Statistics suggest that out of total diabetes patients approximately 21% of patients have DR that is diagnosed in early stage. High sugar and blood glucose level persist in patients for long time the probability of DR is increased [6]. Its major symptoms include primary abnormalities or infections occurring in eyes like focal dilatation of retinal capillaries, which appears like tiny and dark red spots in eyes that is known as microaneurysms (MAs), exudates, and hemorrhages. Here, we have taken many different types of retinal images of diabetic retinopathy patient and extract the important features by applying image processing algorithms, like lesion specific (microaneurysms, exudates), the diameter of the optic disk and image level (prescreening, AM/FM, quality assessment). Collectively all the mentioned features are then provided to the machine learning algorithms like a boosted decision tree, neural network [7], support vector machine (SVM) [8, 9], etc., to get information about DR and non-DR patients.

2 Literature Review

In this section, components of feature extraction are explained briefly.

Image-level Components:

1. Quality assessment: In quality assessment (QA), images are classified on the basis of sufficient quality where the box count values of the noticed vessel system are taken as a feature. For vessel segmentation, the method of Hidden Markov Random Fields was used.
2. Prescreening: In this step [10], images are categorized into normal state and abnormal state. If it is normal then it is to be forwarded for further processing. Each image is divided into disjoint regions and a simple texture descriptor (inhomogeneity measure) is extracted for each region. Then, a machine learning classifier is trained to classify the images based on these features.

drksujatha23@gmail.com

3. Multi-scale AM/FM based feature extraction: In this process, information are obtained from the images and green channel of that images are dissolve into many other forms of representation which shows the intensity, texture, and geometry of the structure using signal processing techniques. By applying the machine learning classifier, images are classified on the basis of extracted features.

Lesion-specific components:

1. One of the early symptoms of DR is microaneurysm. Microaneurysm is red dots in the image of the eye. It is similar to the vessel fragments, so its very hard to detect them.
2. Exudate Detection: It is also one of the early symptoms of DR. It takes place when leakage of fat occurred from the blood vessels.

Anatomical components:

1. Macula Detection: In human eyes structure, macula is present in the center of eye, where center is known as the fovea. If lesions are found in the macula then it results in loss of eyesight because it is situated in the center of the retina.
2. Optic disc detection: The area where optic center enters eye is known as an optic disc. If a person is infected then it displaces from its place.

3 Methodology

Machine learning is an advanced statistical technique. It is the field of artificial intelligence that gives the power of learning to computers from input data. And we can say that it is a statistical technique in which result improves with more data. Machine learning methodology is shown in Fig. 1. There are many machine learning algorithms exist. But, we will consider decision tree, and neural network for the optimization of DR.

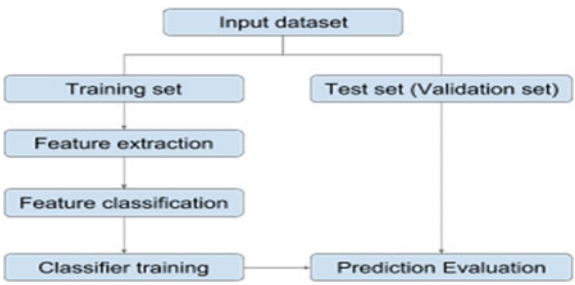


Fig. 1 Machine learning methodology

drksujatha23@gmail.com

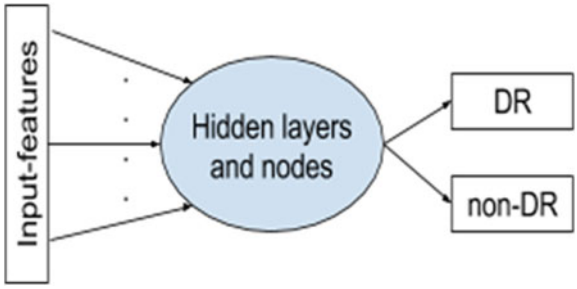


Fig. 2 Flow diagram of neural network

Decision Tree

Decision tree is one of the most widely used machine learning algorithms [11]. This is a nonparametric machine learning algorithm which is used for regression and classification problems. Decision tree learns from simple decision methods inferred from input data features. The main advantages of decision trees include the following:

1. It is very simple and easy to understand and interpret.
2. Decision tree can be visualized.
3. It does not require more data preparation like some other method where we need to normalize data, etc.
4. It uses a very simple model, i.e., if the situation of a given model is known then the condition (or decision) can be easily explained by simple Boolean logic.

Neural Network

Neural networks (NNs) are also called "Artificial Neural Network" or "Deep Learning". It is inspired by the working of brain neurons. The first idea about how neurons might work presented in 1943 by Warren McCulloch, a neurophysiologist and Walter Pitts, a mathematician [12]. It is generally represented by the network diagram as shown in Fig. 2. Neural network applies a transfer function and a threshold which decide how the incoming information will scale and produces an activation function. If the activation function is strong that passes the threshold then it will produce a positive output else negative output. Generally, we take the following function as the activation function:

1. Most popular activation function is sigmoid or a hyperbolic tangent functions, which is a nonlinear function;
2. Threshold binary logic function;
3. Linear function, etc.

drksujatha23@gmail.com

4 Analysis and Result

As we have described in the methodology section, that the first step is to understand the input features. We will first show the statistical results of features and then machine learning algorithm results.

Input Features Statistical Analysis

There are a lot of irrelevant input features that decrease the accuracy of the input model. There are mainly three benefits of removing the irrelevant input.

(a) Decreases chance of overfitting: If there are less redundant data then there is very less probability to make a decision based on the noise.
(b) Accuracy improves: If there are less misleading data then the accuracy automatically increases.
(c) Reduces training and test time: Training time increases linearly with a number of input feature. So, removing irrelevant features will reduce the training and testing timing.

Steps of feature selection:

1. First thing that comes in this category is to view the histogram of input features. Figure 3 shows the comparisons of some important features that distinguish between DR and non-DR patients.
2. **Correlation Matrix**: Figure 4 shows the correlation matrix between each input feature using 3D distribution.

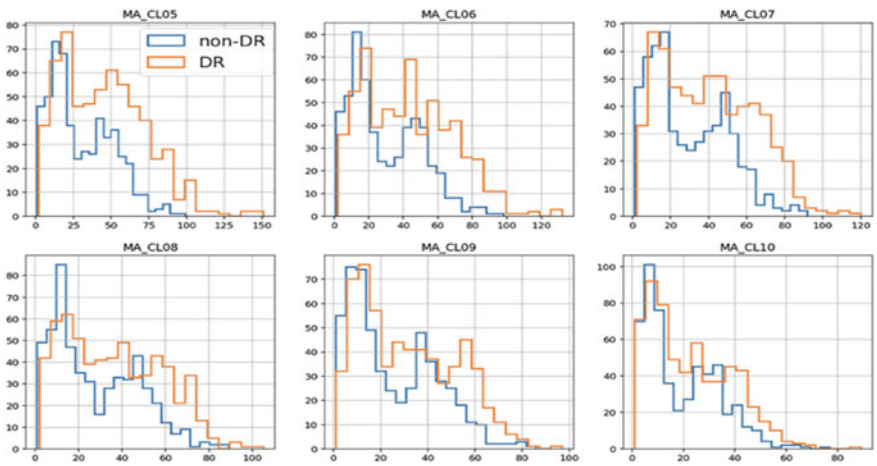


Fig. 3 Comparison of features of DR and non-DR patients

drksujatha23@gmail.com

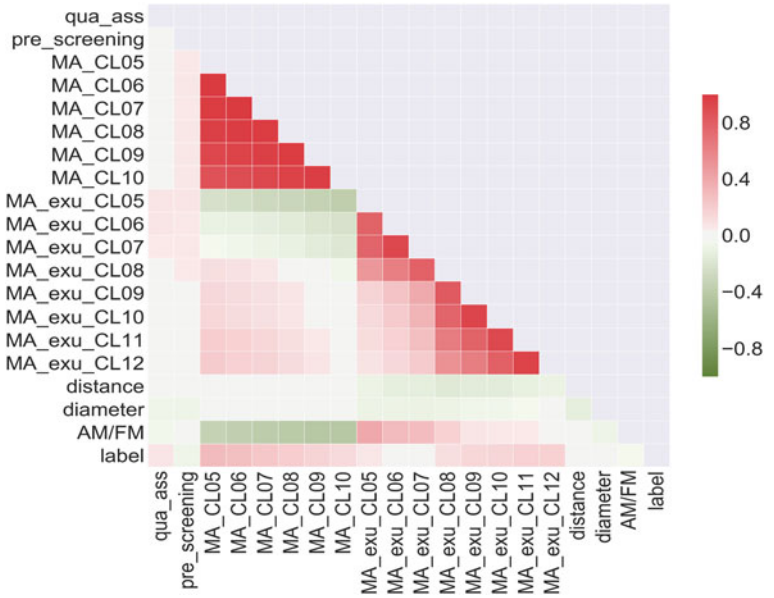


Fig. 4 Correlation matrices

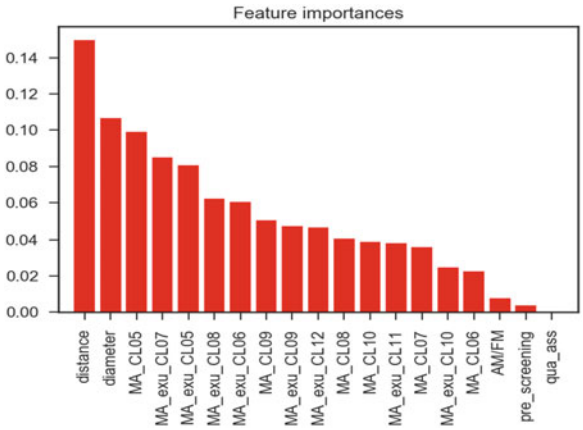


Fig. 5 Feature importance of input variables

3. **Feature Importance**: Checking the relative importance of each input feature. Figure 5, shows the bar plot for the feature importance.

Machine learning results

1. Boosted decision tree:

   (a) Scan of a number of trees with a mean square error. The number of trees increases the mean square error decreases. It is shown in Fig. 6.

drksujatha23@gmail.com

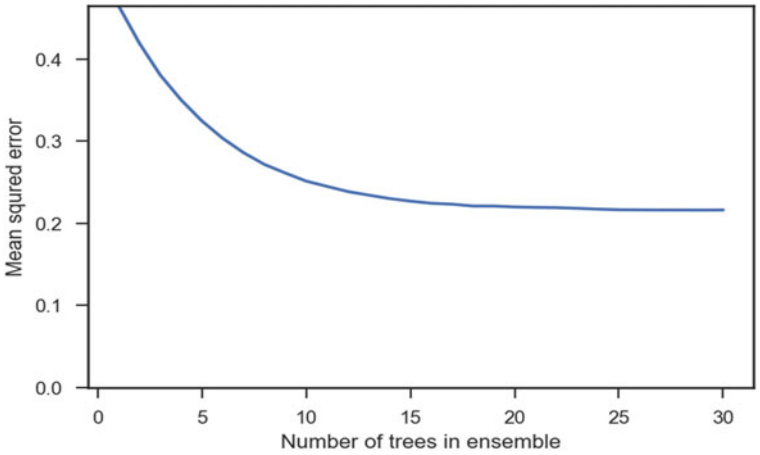


Fig. 6 Number of trees of BDT versus mean square error

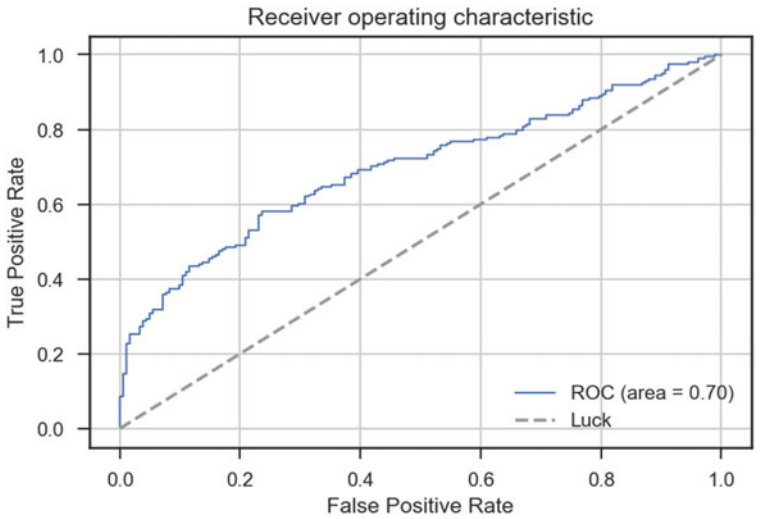


Fig. 7 ROC curve for BDT

(b) Receiver Operating Curve (ROC) Curve: ROC curve is the graphical representation that shows the diagnostic ability of the classifier system as its discriminating threshold is varied. It is shown in Fig. 7.

(c) Overtraining check: It tells us about the learning ability of the model. If our model is not overtrained then the test data and train data will overlap each other. Figure 8 shows the overtraining plot. In this paper, the train and test data is not matching it's because of very low statistics.

drksujatha23@gmail.com

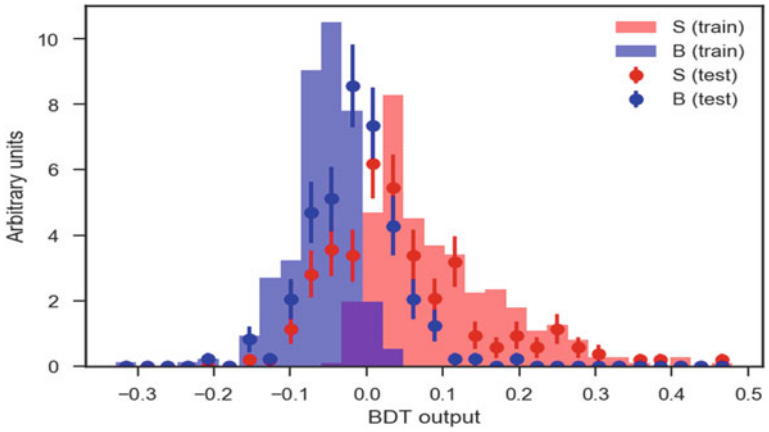


Fig. 8 BDT output of train and test data distribution

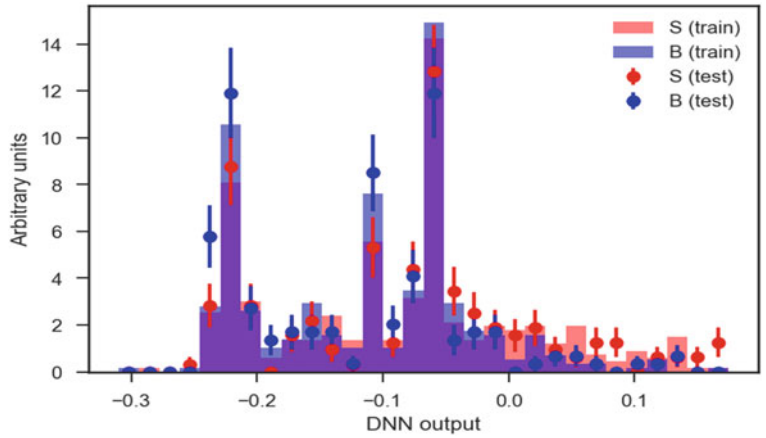


Fig. 9 Neural network output of train and test data distribution

2. Neural Network:

   (a) Overtraining check for the neural network is shown in Fig. 9. Here, the neural network is trained better than BDT.
   (b) Receiver Operating Curve (ROC) Curve for the neural network is shown in Fig. 10.

drksujatha23@gmail.com

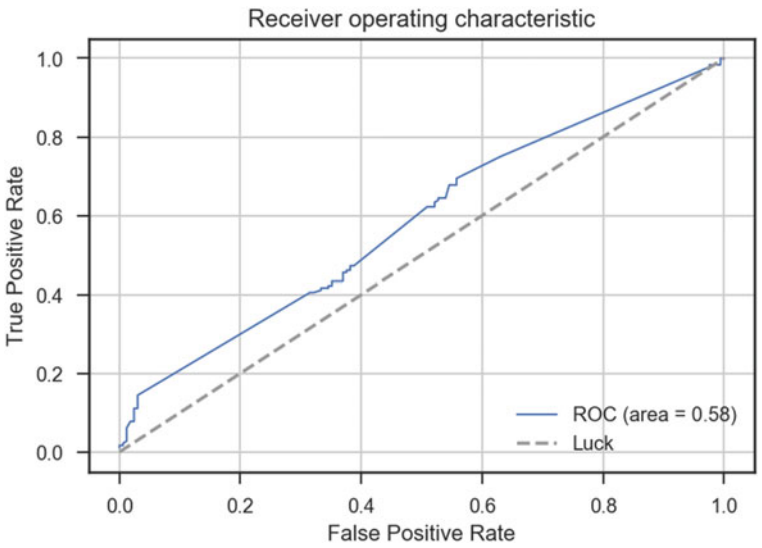


Fig. 10 ROC curve for NN versus mean square error

5 Summary and Conclusion

Disease detection is itself a very challenging task with the use of machine learning techniques, the diagnosis becomes effective as well as convenient. Diabetic Retinopathy (DR) is a decisive disease and spreading all over the world. The experimental work is totally based on open-source technology. In this paper, we used BDT and DNN for analysis and their performance has been shown in the ROC curve. Overtraining is also checked. We are struggling with low statistics that is shown in the overtraining plot of BDT and DNN. So, results are reasonably good considering the low statistics.

References

1. Wild S, Roglic G, Green A, Sicree R, King H (2004) Global prevalence of diabetes: estimates for the year 2000 and projections for 2030. Diabetes Care 27:1047–1053
2. http://www.diabetesforecast.org/diabetes-101/complications.html
3. Diabetic retinopathy. Diabetes.co.uk. Retrieved 25 Nov 2012
4. Prevention of blindness from diabetic retinopathy. Report of a WHO Consultation, Geneva; November, 2005
5. Gadkari SS, Maskati QB, Nayak BK (2016) Prevalence of diabetic retinopathy in India: the all india ophthalmological society diabetic retinopathy eye screening study 2014. Indian J Ophthalmol 64:38–44
6. Fong DS, Aiello L, Gardner TW, King GL, Blankenship G, Cavallerano JD et al (2003) Diabetic retinopathy. Diabetes Care 26(1):s99–s102

drksujatha23@gmail.com

7. Zhang Z (2016) Introduction to machine learning: k-nearest neighbors. Ann Transl Med 4 (11):218. https://doi.org/10.21037/atm.2016.03.37
8. Antal B, Hajdu A (2014) An ensemble-based system for automatic screening of diabetic retinopathy. Knowl Syst 60:20–27. https://doi.org/10.1016/j.knosys.2013.12.023
9. Cortes C, Vapnik V (1995) Mach Learn 20:273. https://doi.org/10.1007/BF00994018
10. Antal B, Hajdu A, Szab-Maros Z, Trk Z, Csutak A, Pet T (2012) A two-phase decision support framework for the automatic screening of digital fundus images. J Comput Sci 3:262–268
11. Quinlan J (1986) Introduction to decision tree. Mach Learn 1(1):81–106. https://doi.org/10.1023/A:1022643204877
12. McCulloch WS, Pitts (1943) A logical calculus of the ideas immanent in nervous activity. Bull Math Biophys 5(4):115–133. https://doi.org/10.1007/BF02478259

drksujatha23@gmail.com

Diagnosis of Parkinson's Diseases Using Classification Based on Voice Recordings



P. Chitra Rajagopal, Tanupriya Choudhury, Archana Sharma and Praveen Kumar

Abstract Machine learning techniques prove to be very efficient when it comes to classifying things. Currently, most of the health practices rely on the opinions of the clinicians for making the correct diagnosis. This makes it difficult for people who cannot afford to go to the specialist due to the shortage of funds and also there are many cases where the disease goes undetected for a long period of time and thus leading to a very low survival rate for the patient. Many advances in technology have been made in order to reduce the errors in diagnosis and thus, reaching a better conclusion faster. This paper aims at bringing those developments to the light. This all is done in hope of getting a fair idea of the present situation and thus forming a better plan of how we should all work toward achieving better results. Also, a proposal is made, comparisons of which can be done with the existing algorithms in order to make a contribution in the field of diagnosis of diseases.

Keywords Machine learning · Neural networks · Disease diagnosis

1 Introduction

Parkinson's disease is a progressive neurodegenerative disease that attacks a particular type of neurons in the brain, the dopamine-producing neurons. Substantia Nigra, the house of dopamine-producing neurons is affected in this disease. In India

P. Chitra Rajagopal · T. Choudhury (✉) · A. Sharma · P. Kumar
Computer Science and Engineering Department, Amity University Uttar Pradesh, Noida, Uttar Pardesh, India
e-mail: tchoudhury@amity.edu; tanupriya1986@gmail.com

P. Chitra Rajagopal
e-mail: chitrargpnirmal@gmail.com

A. Sharma
e-mail: archanabtech1501@gmail.com

P. Kumar
e-mail: pkumar3@amity.edu

© Springer Nature Singapore Pte Ltd. 2019

V. S. Rathore et al. (eds.), *Emerging Trends in Expert Applications and Security*, Advances in Intelligent Systems and Computing 841,
https://doi.org/10.1007/978-981-13-2285-3_68

drksujatha23@gmail.com

alone, over one million people are diagnosed with PD [Parkinson's disease] every year [1].

Neural networks are inspired by the structure of the human brain. The human brain is capable of performing a very complex set of tasks by using a similar set of structures [2]. Inspired by such an efficient system of problem solving, neural networks pave the way for human beings to build a machine which can teach itself how to solve problems without having to be programmed specifically to do so [3].

The weights are adjusted in order to generate an output which is closer to the desired output. The error is calculated for each case as follows:

$$\text{Error} = \text{Desired output} - \text{Generated output}$$

$$\Delta_i = T_i - O_i * g'\left(\sum_j W_{j,i}, a_j\right)$$

$$W_{j,I} \leftarrow W_{j,i} + \alpha * a_j * \Delta_i$$

The goal is to reduce the error as much as possible and this is done by adjusting the weights by training the neural network by feeding it with a plethora of learning cases. This makes neural networks efficient but slow learners.

2 Problem Definition

Since there is no reliable known cause for PD, it becomes difficult to catch the disease at an early stage until any of the visual symptoms are evident. Moreover, many of the patients come to the clinicians who prefer MRI scans and other expensive methods to make a diagnosis and set a treatment course. Not only it is an expensive method but also it provides little relief to people who reach a wider set of visible symptoms.

In such a scenario, we need to employ a method which can not only detect PD but also be able to do that at a reduced cost. This is where voice analysis comes into question. Human ear might not be able to differentiate between the voices of healthy individuals from those of the ones with PD but a computer can be trained to do so. A neural network can be fed with data that teaches it to extract the differences between these two sets of the population.

3 Architecture and Model

This research has been carried out on the dataset provided by the UCI machine learning repository [4]. The particular dataset utilized in this paper contains biomedical voice measurements from 31 people, out of which 23 individuals are diagnosed with PD. A set of 23 attributes that have been evaluated by the neural

drksujatha23@gmail.com

Table 1 Choice of neurons in the hidden layer

Sr. no.	Epochs	Number of neurons in the hidden layer	Accuracy
1.	500	11	93.33
2.		12	99.49
3.		13	94.36
4.		14	95.38
5.		15	94.87
6.	400	11	94.87
7.		12	97.44
8.		13	94.87
9.		14	89.74

network model built for this purpose all of those are enlisted in Table 1. For every individual (whether healthy or having PD), 6 voice recordings were taken giving a total of 195 voice recordings. This dataset is fed to a neural network which uses this training set to adjust the weights for different values so that maximum efficiency can be obtained. The neural network model consists of three layers; input layer, hidden layer, and the output layer. Since there are 23 attributes, therefore, the input layer consists of 23 neurons. The output layer has to provide an output in the form of zeroes and ones; hence, it makes use of the sigmoid function to do so. The number of neurons to be put inside the hidden layer varies from problem to problem, that is, a particular number of neurons might lead to great results in a peculiar scenario but it may not in a completely different situation. In this particular classification problem, the following formula has been used to calculate the number of neurons in the hidden layer of the feed-forward neural network:

$$N_h = (N_i + N_o)/2$$

where N_i is the number of neurons in the input layer and N_o is the number of neurons in the output layer.

Here, from Fig. 1, it is clearly evident from the bar graph that the exact relationship between the number of neurons in the hidden layer and the accuracy achieved by the model is somewhat difficult to determine. It is only by hit and trial in this case that it is found that the best choice for the number of neurons in the hidden layer is obtained by calculating the mean of the number of neurons in the input and the output layer.

In other cases, some other number might generate the desired results [5].

Coming to the activation functions used in different layers, it suffices to say that a combination of Rectified Linear Unit and Sigmoid functions served the purpose well. The working of these activation functions is explained below.

drksujatha23@gmail.com

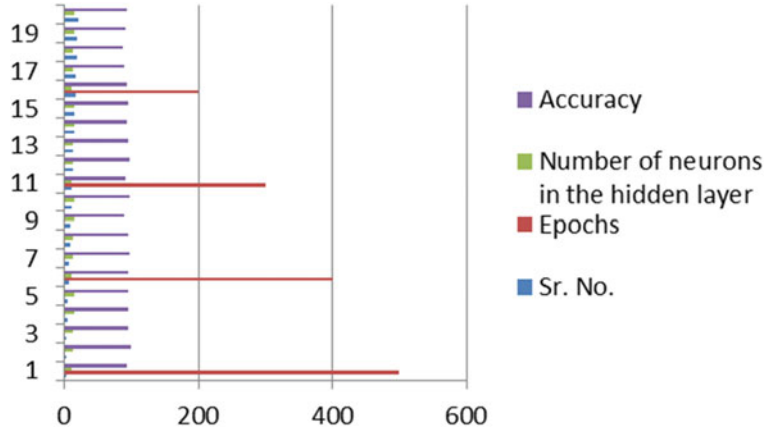


Fig. 1 Graph showing the accuracy achieved at different epochs and the number of neurons

Sigmoid: The mathematical form of the sigmoid function is as follows:

$$F(x) = 1/(1 + e^{-x})$$

Sigmoid function takes in a real value and then jams it into a range between 0 and 1. Numbers falling on the side of large negative numbers come out to be zero and numbers falling on the side of large positive numbers come out to be one. It has been used in this paper due to two main reasons:

1. It provides a very simple yet efficient way of binary classification because there are only two outputs: 0 and 1.
2. It provides a nice understanding of the firing rate of a neuron, that is, there are just two cases: either it influences other neurons or it does not.

Rectified Linear Unit: The mathematical form of the rectified linear unit is as follows.

$$F(x) = \max(0, x)$$

Table 2 Advantages of the functions used

Sr. no.	Function used	Advantages (specific to this case)
1.	Sigmoid	• Provides a way of classifying healthy and sick individuals by 0 and 1 by simple association with 0 and 1 • Shows the influence of one neuron over others in a simple manner
2.	Rectified linear unit	• It has a faster convergence rate • It uses less expensive operations

drksujatha23@gmail.com

It is evident from the function definition that the activation threshold is simply achieved at zero. This function provides the following advantages:

1. The convergence rate of gradient descent is much faster in this function as compared to sigmoid and tanh functions.
2. It achieves almost the same set of results with the help of relatively less expensive operations (Table 2).

3.1 Learning Process and Training

Before actually feeding data to a model, we need to provide it with a set of instructions on how it should learn from the basic data it is fed with. A loss function should be clearly defined. The goal is to minimize the loss function. It tells the system how it should adjust its weights in order to minimize the difference between the target output and the actual output.

After directing the system in a way it should work, the next step comes to training it by showing it a number of examples with the help of which it can make a number of informed choices itself as to how it should adjust the weights in order to minimize the losses. The model should fit itself in a way as to cover the maximum number of points provided in the training sample.

4 Results

As it has been stated before that the training cases decide the way a system learns but there are many other factors, which are responsible for making the system ready to make decisions over unseen sets of data. The variables on which the accuracy depend on are the following:

1. Epochs: It is the variable which decides how many times all the training cases are to be tried before the weights to minimize the losses are updated.
2. Batch size: This variable decides the number of training cases which are processed before updating the weights once.

Table 3 Accuracy observed with a different set of variables

Sr. no.	Epochs	Batch size	Accuracy (%)
1.	500	6	99.49
2.	500	10	96.59
3.	475	6	98.46
4.	475	10	95.90
5.	450	6	97.95

drksujatha23@gmail.com

3. Number of neurons in the hidden layer: The choice of neurons in the hidden layer plays a crucial role in determining the accuracy achieved by the model as is evident from Table 3 and Fig. 1.

Therefore, in this experiment, a number of variables were tried and the maximum efficiency was observed for a particular set of variables as shown below.

Figure 2 shows how the accuracy of the neural network model stays almost the same even over a wide range of values for the given variables (Fig. 3).

Also, the loss and accuracy for several cases were calculated and it was observed that the no matter what values of variables are used, the loss always deprecates and the accuracy always increases.

Future proposal and conclusion

It is evident from the results that even a simple model of feed-forward neural network can generate a system of classification which is quite efficient to be put into practical use [6]. The accuracy is more than 99% for a particular set of variables. Although one cannot count on this method for making the diagnosis entirely but still, it can be extremely beneficial not only for the clinicians but also for the patients who may not always be able to afford the expensive methods like Magnetic

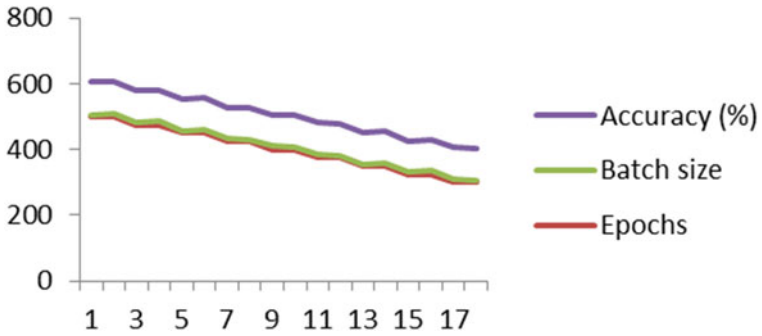


Fig. 2 Relationship between batch size, epochs, and accuracy

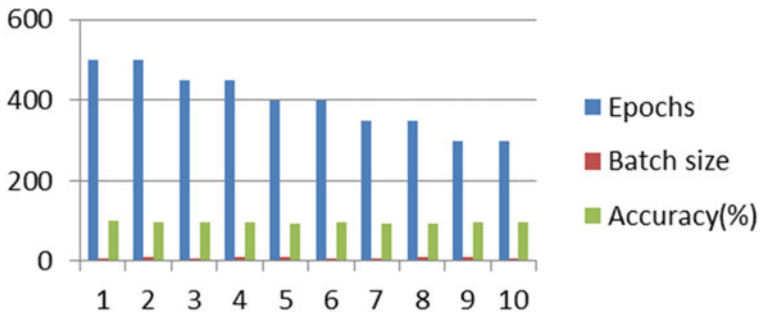


Fig. 3 Accuracy over a range of epochs and batch sizes

drksujatha23@gmail.com

Resonance Imaging to embark on the course of treatment. In this paper, it is shown that how neural networks can be put into use to obtain effective results which may, in turn, be used to cure a very commonly incurred progressive disease of the brain.

After all, the secret to curing the disease of the brain lies somewhere in the brain itself. (Where the inspiration to build something like neural networks come from).

References

1. Google search (Source Apollo Hospital). https://g.co/kgs/4zwICd
2. Schmidhuber J (2014) Deep learning in neural networks: an overview. Elsevier
3. Maind SB, Wankar P (2014) Research paper on basic of artificial neural network. Int J Recent Innov Trends Comput Commun 2(1)
4. Little MA, Costello DAE, Roberts SJ, McSharry PE, Moroz IM (2007) Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection. BioMed Eng OnLine, 1–4
5. Ahamed K, Akhtar S (2016) Survey on artificial neural network learning technique algorithms. Int Res J Eng Technol (IRJET) 3(2):1–5
6. More S, Dhir SG, Daiwadney D, Dhir SR (2016) Review on natural language translator using quantum neural network (QNN). Int J Eng Techn 2(1):1–5

drksujatha23@gmail.com

Analytical Analysis of Learners' Dropout Rate with Data Mining Techniques



Shivanshi Goel, A. Sai Sabitha, Tanupriya Choudhury and Inderpal Singh Mehta

Abstract Massive open online courses (MOOC) are the handy ways for offering the access to quality education especially to those who are pursuing distant education or who wants to enlighten the core of course from the brilliant tutors. The student population has a tendency to be youthful, knowledgeable and MOOC helps students to be a part of different learning methodologies. The basic objective of this work is to understand MOOC environment and analyse the issues related to these courses. The scope of this research work is to understand the major issue of dropout in MOOC courses. Data mining techniques are used to predict this factor which leads to an increase in dropout rates. By identifying and understanding these factors, necessary measures can be taken to push similar types of learners to their maximum potential in the subsequent MOOC courses. This eventually leads to an increase in the completion rate of the MOOC learners.

Keywords MOOC · Data mining · SPSS · Chi-square test · Logistic regression

1 Introduction

MOOC courses target unlimited participation and provide open online training worldwide through the Internet. This course provides us with advantages like communicating and participating with intellectuals' of the same group. According to Yousef et al. [1] these courses offer a very flexible environment thus learning can

S. Goel · A. Sai Sabitha · T. Choudhury (✉) · I. S. Mehta
Amity University Uttar Pradesh, Noida, Uttar Pradesh, India
e-mail: tchoudhury@amity.edu; tanupriya1986@gmail.com

S. Goel
e-mail: Shivanshigoel.27@gmail.com

A. Sai Sabitha
e-mail: saisabitha@gmail.com

I. S. Mehta
e-mail: inderpalmehta@gmail.com

© Springer Nature Singapore Pte Ltd. 2019

V. S. Rathore et al. (eds.), *Emerging Trends in Expert Applications and Security*,
Advances in Intelligent Systems and Computing 841,
https://doi.org/10.1007/978-981-13-2285-3_69

drksujatha23@gmail.com

be done anywhere and anytime. They increase the interest in learners as they provide short videos, quizzes during the course of learning. David George Glance et al. [2] suggested the MOOC characteristics' features are: online method of gaining knowledge, online tests and appraisals, short recordings, online discussions, openness and barriers to persistence. MOOC's online test makes the topic clear as tests are given just after the completion of the topic. Learning through short recording makes the topic clear and they do not bore or distract the users from their topic of interest. Sir Daniel [3] depicted MOOC as "the instructive popular expression of 2012". Siemens [4] and Pappano [5] also said that MOOC is a popular learning platform and a huge number of individuals have enrolled in these courses. DeWaard et al. [6] features about MOOC's are openness, self-organization and connectedness. As indicated by the New York Times, MOOCs are a piece of a sensational movement in online instruction and they can possibly reform how formalized instruction is conveyed. The objective of this research paper is to review the MOOC design and to understand the important issues that are affecting the course. A case study was conducted to identify the factors affected the dropout rate. The paper is structured as follows; Sect. 2 is a literature survey, Sect. 3 is method and evaluation of MOOC and data mining, Sect. 4 is the case study, Sect. 5 is results of case result and future scope.

2 Literature Survey

2.1 MOOC Design

Brown and Voltz [7] gave six elements for effective MOOC design which are action, situation, response, delivery, context, an influence which helps with the route of the complexities which impact the improvement of a successful MOOC outline. Casazza [8] said that distinct frameworks of communication appear to be at the heart of a large portion of the social and ethnic contrasts that influence the learning environment. Jung [9] suggested the analytics steps into requirements analysis, content research, innovative and natural examination. Tanmay Sinha [10] suggested two imperative encouraging factors which are behaviour activities and metric of interest. Siemens [11] classified MOOC as in x-MOOC, c-MOOC and quasi-MOOC the class of quasi MOOCs incorporates online instructional exercises as OER that are in fact not courses, however, are expected to bolster learning-particular undertakings and comprise of non-concurrent learning assets. Gayoung LEE [12] constructs a design model which consists of six stages: analysis phase1, design, development, implementation, evaluation, analysis phase2. This model has two literature reviews and expert reviews for external justification can be used for testing usability. Discussions in MOOCs have a tendency to draw in just a

drksujatha23@gmail.com

Table 1 Consist of author names with their Frameworks and Findings

Author names	Frameworks	Findings
Daradoumis et al. [13]	Agent-based framework	Framework for improving delivery, efficiency
Gynther [14]	Designing framework	Teachers, participants and contents
Alario-Hoyos [15]	Conceptual framework	Learners, contents, complementary technologies
Gulati [16]	Cognitive agent based framework	record of substance and Altered improvement
Dalziel [17]	Broad-reaching Learning design	Client organization and instructor creating/ adjustment of successions.
Osterwalder and Pigneur [18]	MOOC canvas design	Human, intellectual, equipment and platform
Lim and Kim [19]	SWOT analysis	Type of organization, learners age, lecture areas, system connection, etc.

little parcel of the understudy movement Jacqueline Aundree Baxter and Jo Haycock [20]. This is setting discussions in MOOCs separated from 'instructional exercise sort' gatherings used to bolster understudies' learning in online or mixed courses in higher instruction. Besides, some contend that dynamic engagement is definitely not the main method for profiting from discourse gatherings. Kim and Lim [19] and understudies' qualities and inclinations could be more vital than the course outline in deciding the route in which they take the full favourable position of online assets [13]. It recommended outline steps as substance configuration, association outline and standard outline and some of them are following (Refer Table 1).

2.2 Factors Influencing MOOC

Su white [21] identifies that knowledge enrichment, increase in confidence level and practical-oriented teaching–learning process are the motivating factors which make MOOC course interesting to the learners'. Zhu [22] finds that learners' performance is correlated with variables like instructor, course, course year, course size, staff size, content on the course websites and the existence of forums on the course website. Marcus Klüsener (2015) used data mining techniques to understand the success rate of students and he concluded that academic performance, persistence, retention, program satisfaction are the factors for success or course completion. (Kim [19]) gave connecting achievement with tirelessness focused on learner utilization and cooperation with course segments, recordings and exchange gatherings. From the above study, it was found that knowledge enrichment due to

drksujatha23@gmail.com

personalized course content, face-to-face instructional delivery and discussion forums are some of the factors that motivate students to enroll in these courses. Though MOOC has various advantages, we study and show that there are some issues which have to be conducted. To understand the various issues in MOOC the further literature study was conducted and given below.

Dong Clow [23] gave four stages to understand the drop-out rate which are: Awareness, Registration, Activity and Progress. Xiaohong Su [24] gave the factors that obstruct students' learning which are supervision, lack of a goal, lack of environment. Tinto et al. [25] explain students' dropout in as socio-mental procedures that happen as students' move from their life preceding college cooperation and their new engagement in college life. Zhu [22] suggests no commitment to complete the course is one of the factors which increase the dropout rates. According to the above study, the dropout rate of learners' is an important issue in MOOC courses.

The basic factors for content quality in a MOOC environment are objective, understanding ability, visibility, accessibility, language, evaluation and privacy. Siemens et al. [4] [11] suggested two criteria to analyse the content of the course which are automated analysis whose factors are key-concept, clustering and in-depth analysis and manual analysis factors are qualitative, quantitative. According to Bauer [26], content investigation is a deliberate strategy for coding the content and pictures. McMillan [27] provided techniques which analyse the content of course by articulating research questions, checking of unwavering quality of code, selection of computer-mediated data, discourse features, etc. Mayring [28] and Katrien Verbert suggested some content models which are learnability content model, SCORM content aggregation model, reusable learning object, and NETg learning model. The above study helped in understanding the important factors responsible for content quality. Thus dropout rates, quality of content and content delivery to the learners' are important issues in MOOC. For analysing the dropout rate, following experimental work is done.

3 Methods and Evaluations of MOOC and Data Mining

Data mining is used to mine the information from a data set and renovate into a logical structure for further use which involves data preprocessing, model and inference considerations, visualization, etc. Some of the techniques used in data mining are classification, clustering, decision tree, etc. In classification is the commission of generalizing known structure to apply to new data. Whereas clustering is the task of realizing groups and structures in the data that are in some way or another "similar", without using known structures in data and, in decision tree, the root of the tree is a simple question or condition that has multiple answers. The classification has a sub-technique that is logistic regression which is a regression model used to estimate the probability of a binary based on independent and dependent variables whereas dependent variables can be in the form of the

drksujatha23@gmail.com

Table 2 DM Techniques and MOOC

Author names	Findings	Techniques used	Year
Betanzos Atienza, M.	Human translation quality in MOOC environment	SVM	2015
Glassman, E. L.	Interacting with massive numbers of student solutions	Clustering	2014
Colin Taylor et al.	Predicting Stop out in MOOC's	Logistic regression	2014
Christopher, G. Brinton et al.	Video-watching behaviour versus video quiz performance	SVM	2016
Jiezhong Qiuynd al.	Learning behaviour in MOOCs	Logistic Regression, SVM	2016
Suhang Jiang et al.	MOOC performance	Logistic regression	2014
Gregor Kennedy et al.	Predict MOOC performance	Multiple regression	2015
Vitomir Kovanovićnd al.	Student behaviour change between two course enrolments	Clustering	2016
Miaomiao Wen et al.	Discussion forums	Sentiment analysis	2014
Aneesha Bakharia	Forum post	Naïve Bayes, SVM	2016
Adamopoulos, P.	Student retention	Sentiment analysis	2013
Sinha, T.	Information processing and attrition behaviour	Markov model	2014

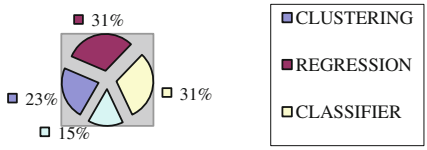


Fig. 1 Graphical representation of data mining techniques used in MOOCs

categorical variable (yes/no) or binary variable (0/1) Its outcome is given by a specific percentage. Data mining techniques can transform the learning processes and techniques like classification, prediction, association and clustering are used in an e-learning environment. The other data mining techniques that are used in different areas of MOOC are shown in the table given below (Refer Table 2) and (Fig. 1).

drksujatha23@gmail.com

4 Result and Analysis

Data gathering/collection: The dataset used for this study is the HarvardX-MITx Person-Course de-identified dataset AY2013 from (https://dataverse.harvard.edu/dataverse/mxhx) released on May 27, 2014. It contains 641,138 rows and 14 columns. The dataset depicts learner's engagement in courses one of the leading MOOCs platforms. Each row represents a learner who is enrolled in any one of the courses in MITx or Harvardx offered on edX platform for the academic year 2013. Screenshot of the dataset (Refer Fig. 2).

Methodology: To understand the factors that influence the dropout rate, chi-square test and logistic regression are applied to test the data. Chi-square test is most commonly used to evaluate test of independence when using a cross-tabulation which introduces the conveyances of two clear-cut factors, with the convergences of the classes of the factors. Figuring chi-square test and looking at it against a critical value from the chi-square distribution permits the researchers to survey whether the association seen between the variables in a particular sample is to represent the actual relationship between those variables. The dependent variables considered for logistic regression given below. The target variables considered for logistic regression are certified and performance is given below. The experiment is conducted using SPSS. The dependent_variable and target_variable dataset are given below.

Dependent variable: n_days_act, n_events, n_play_video, n_chapters, n_forum_posts and Target Variable: Certified, Incomplete_flag.

Chi-Square Test:

1. It is used for finding those factors due to which learners' dropout rate increases.
2. Research hypothesis (h1): It proposes that two variables are dependent on each other.
3. Null hypothesis (h0): It proposes that two variables are independent of each other.
4. If the value of p < chi-square value then we reject the null hypothesis and if the value of p > chi-square value then we reject the research hypothesis (Table 3).

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	course_id	userid_DI	registered	viewed	explored	certified	final_cc_cname	LoE_DI	YoB	gender	grade	start_time_DI	last_event_DI	nevents	ndays_a	nplay_vi	nchapter	nforum_p	roles	incomplete_flag	
2	HarvardX/CB22x/2	MHvPC1304426	1	0	0	0	United States	NA	NA	NA	0	19-12-2012	17-11-2013		9			0		1	
3	HarvardX/CS50x/2	MHvPC1304426	1	1	0	0	United States	NA	NA	NA	0	15-10-2012			9		1	0		1	
4	HarvardX/CB22x/2	MHvPC1302758	1	0	0	0	United States	NA	NA	NA	0	08-02-2013	17-11-2013		16			0		1	
5	HarvardX/CS50x/2	MHvPC1302758	1	0	0	0	United States	NA	NA	NA	0	17-09-2012			16			0		1	
6	HarvardX/ER22x/2	MHvPC1302758	1	0	0	0	United States	NA	NA	NA	0	19-12-2012			16			0		1	
7	HarvardX/PH207x/	MHvPC1302758	1	1	1	0	United States	NA	NA	NA	0	17-09-2012	23-05-2013	502	16	50	12	0			
8	HarvardX/PH278x/	MHvPC1302758	1	0	0	0	United States	NA	NA	NA	0	08-02-2013			16			0		1	
9	HarvardX/CB22x/2	MHvPC1305394	1	1	0	0	France	NA	NA	NA	0	01-01-2013	14-05-2013	42	6		3	0			
10	HarvardX/CB22x/2	MHvPC1300883	1	1	0	0	United States	NA	NA	NA	0	18-02-2013	17-03-2013	70	3		3	0			
11															12			0		1	

Fig. 2 Screenshot of the dataset used in this paper

drksujatha23@gmail.com

Table 3 Result of chi-square test

Factors affecting dropout rates	Chi-square tests	Summary
N_EVENTS	Pearson Chi-Square: Value 4896.098[a], df 797, Asymp. Sig. (2-sided) .000 Likelihood Ratio: 6511.111, 797, .000 Linear-by-Linear Association: 167.681, 1, .000 N of Valid Cases: 14300 a. 1438 cells (90.1%) have expected count less than 5. The minimum expected count is .30.	P = 0.09 and chi-square value is 3.84. So, 0.09 < 3.84 it means that rejection of the null hypothesis and it clearly shows that incomplete_flag is dependent on n_events
N_FORUM POST	Pearson Chi-Square: Value 84.549[a], df 5, Asymp. Sig. (2-sided) .000 Likelihood Ratio: 139.558, 5, .000 Linear-by-Linear Association: 62.199, 1, .000 N of Valid Cases: 14300 a. 4 cells (33.3%) have expected count less than 5. The minimum expected count is 1.52.	P = 0.033 and chi-square value is 3.84. 0.033 < 3.84, which means that incomplete_flag is dependent on nforum_posts
N_DAYS ACT	Pearson Chi-Square: Value 1253.574[a], df 92, Asymp. Sig. (2-sided) .000 Likelihood Ratio: 1326.951, 92, .000 Linear-by-Linear Association: 184.720, 1, .000 N of Valid Cases: 14300 a. 116 cells (62.4%) have expected count less than 5. The minimum expected count is .30.	P = 0.062 and chi-square value is 3.84 0.062 < 3.84, which means the incomplete_flag is dependent on ndays_act
N_PLAYS VIDEOS	Pearson Chi-Square: Value 342.778[a], df 211, Asymp. Sig. (2-sided) .000 Likelihood Ratio: 556.887, 211, .000 Linear-by-Linear Association: 35.737, 1, .000 N of Valid Cases: 14300 a. 396 cells (93.4%) have expected count less than 5. The minimum expected count is .30.	P = 0.093 and chi-square value is 3.84 0.093 < 3.84, means incomplete_flag dependent on nplay_video
N_CHAPTERS	Pearson Chi-Square: Value 2611.558[a], df 34, Asymp. Sig. (2-sided) .000 Likelihood Ratio: 2628.320, 34, .000 Linear-by-Linear Association: 98.489, 1, .000 N of Valid Cases: 14300 a. 27 cells (38.6%) have expected count less than 5. The minimum expected count is .30.	P = 0.038 and chi-square value is 3.84 0.038 < 3.84 which shows that incomplete_flag is dependent on N chapters

From the table (Refer Fig. 3), it is found that incomplete_flag is dependent on variables that are n_days_act, n_events, n_play_video, n_chapters and n_forum_posts. Further logistic regression is done to test the accuracy of dropout rate in the MOOC courses.

Logistic regression Logistic regression is used to classify the variable which targets in the class label that is certified and incomplete_flag. The result of the classification (Refer Fig. 3). 98.2% accuracy on data (1 iteration) was found. The process was further repeated (8 iterations) and it was found the accuracy improved with 98.7% (Refer Fig. 4). The case study concluded that incomplete_flag is depended on n_days_act, n_events, n_play_video, n_chapters and n_forum_posts and it was found that incomplete_flag is observed with an accuracy of 98.7%.

drksujatha23@gmail.com

Classification Table[a,b]

Observed			Predicted		
			certified		Percentage Correct
			0	1	
Step 0	certified	0	14040	0	100.0
		1	260	0	.0
	Overall Percentage				98.2

a. Constant is included in the model.

b. The cut value is .500

Fig. 3 Screenshot of result which finds out the number of certified and non-certified learners

Contingency Table for Hosmer and Lemeshow Test

		certified = 0		certified = 1		
		Observed	Expected	Observed	Expected	Total
Step 1	1	2139	2139.000	0	.000	2139
	2	1467	1467.000	0	.000	1467
	3	748	747.986	0	.014	748
	4	3669	3660.656	0	8.344	3669
	5	1430	1426.362	0	3.638	1430
	6	1431	1426.364	0	4.636	1431
	7	1429	1423.616	1	6.384	1430
	8	1727	1749.017	259	236.983	1986

Classification Table[a]

Observed			Predicted		
			certified		Percentage Correct
			0	1	
Step 1	certified	0	13978	62	99.6
		1	126	134	51.5
	Overall Percentage				98.7

a. The cut value is .500

Fig. 4 Screenshot of the result of iteration table for incomplete_flag

5 Conclusion and Future Scope

This paper discusses the various MOOC concepts and issues of MOOC. An extensive literature study was conducted and from this study dropout rate and content design were identified as the issues in MOOC. Logistic regression and chi-square test were used and the results showed that n_days_act, n_events, n_play_video, n_chapters and n_forum_posts are the factors contributing to the dropout which eventually led to poor results. The dataset considered for the above research work was HarvardX-MITx MOOC-Course. The future research work can be done to understand whether these parameters cohort with other course design. Further, the work can be extended to understand and enhance content quality using data mining techniques.

drksujatha23@gmail.com

References

1. Yousef AMF, Chatti MA, Schroeder U, Wosnitza M, Jakobs H (2014) MOOCs–a review of the state-of-the-art. In: Proceedings of the CSEDU 2014 conference, vol 3, pp 9–20. INSTICC
2. Glance DG, Forses M, Riley M (2013, May). The pedagogical foundations of massive open online courses. http://firstmonday.org/article/view/4350/3673#author
3. Daniel J (2012) Making sense of MOOCs: musings in a maze of myth, paradox and possibility. J Inter Media Edu 18:1–20
4. Siemens G (2012) MOOCs are really a platform. Elearnspace. Retrieved from: http://www.elearnspace.org/blog/2012/07/25/moocs-are-really-a-platform/
5. Pappano L (2012, November 2). The year of the MOOC. The New York Times. http://goo.gl/6QUBEK
6. DeWaard I, Abajian S, Gallagher MS, Hogue R, Keskin N, Koutropoulos A, Rodrigu ez OC (2011) Using mLearning and MOOCs to understand chaos, emergence, and complexity in education. Int Rev Res Open Distance Learn 12(7):94–115
7. Brown AR, Voltz BD (2005) Elements of effective e-learning design. Int Rev Res Open Distrib Learn 6(1)
8. Casazza ME, Silverman SL (2000) Learning and development: making connections to enhance teaching. Jossey-Bass, San Francisco
9. Jung IS (1997) Network-based ISD model for distance learning design of corporate education, a way for the 21st Korean corporate education. J Korean Soc Learn Perform, 44–63. http://www.thearling.com/text/dmwhite/dmwhite.htm
10. Sinha T (2014b) Supporting MOOC instruction with social network analysis. arXiv, preprint arXiv:1401.5175
11. Siemens G (2013) Massive open online courses: innovation in education? In: Commonwealth of learning, perspectives on open and distance learning: open educational resources: innovation, research and practice, p 5. Retrieved from http://www.col.org/resources/publications/Pages/detail.aspx?PID=44 6 on 21/5/13
12. Gayoung LEE, Sunyoung KEUM, Myungsun KIM, Yoomi CHOI, Ilju RHA (2016) A Study on the development of an MOOC design model. Edu Technol Int 17(1):1-37
13. Daradoumis T et al (2013) A review on massive e-learning (MOOC) design, delivery and assessment. In: 2013 eighth international conference on P2P, parallel, grid, cloud and internet computing (3PGCIC). IEEE
14. Gynther K (2016) Design framework for an adaptive MOOC enhanced by blended learning: supplementary training and personalized learning for teacher professional development. Electron J e-Learn 14(1)
15. Alario-Hoyos C, Pérez-Sanagustín M, Cormier D, Kloos CD (2014) Proposal for a conceptual framework for educators to describe and design MOOCs. J UCS 20(1):6–23
16. Gulati N (2013, December) Framework for cognitive agent-based expert system for metacognitive and collaborative E-Learning. In: 2013 IEEE international conference on MOOC innovation and technology in education (MITE), pp 421–426. IEEE
17. Dalziel J (2003). Implementing learning design: The learning activity management system (LAMS)
18. Osterwalder A, Pigneur Y (2010) Business model generation. John Wiley & Sons Inc, Hoboken/New Jersey
19. Lim K, Kim MH (2014) A SWOT analysis of design elements of Korean MOOCs. J Dig Converg 12(6):615–624
20. Jacqueline Aundree Baxter, Jo Haycock (2014) Roles and student identities in online large course forums: Implications for practice. http://www.irrodl.org/index.php/irrodl/article/view/1593/2763
21. White S, Davis H, Dickens K, León M, Sánchez-Vera MM (2014) MOOCs: What motivates the producers and participants?. In Computer Supported Education (pp. 99-114). Springer International Publishing

drksujatha23@gmail.com

22. Zhu K (2014) Research based on data mining of an early warning technology for predicting engineering students' performance. World Trans Eng Technol Edu 12(3):572–575
23. Clow D (2013) MOOCs and the funnel of participation. Proceedings of the 3rd International Conference on Learning Analytics and Knowledge (LAK '13), 8–12 April 2013, Leuven, Belgium (pp. 185–189). New York: ACM
24. Xiaohong Su, Tiantian Wang, Jing Qiu, Lingling Zhao (2015) Motivating students with new mechanisms of online assignments and examination to meet the MOOC challenges for programming. IEEE Frontiers in Education Conference 2015:1–6
25. Tinto V (1975) Dropout from higher education: A theoretical synthesis of recent research. Rev Edu Res 45(1):89-125
26. Bauer M (2000) Classical content analysis: A review. In: Bauer MW, Gaskell G (eds) Qualitative researching with text, image, and sound: a practical handbook. Sage, London, pp 131–151
27. McMillan, SJ (2000) The microscope and the moving target: The challenge of applying content analysis to the World Wide Web. Journalism and Mass Communication Quarterly 77 (1):80–98.
28. Mayring P (2014) Qualitative content analysis: theoretical foundation, basic procedures and software solution

drksujatha23@gmail.com

Discovering the Unknown Patterns of Crop Production Using Clustering Analysis



Dakshita Sharma, A. Sai Sabitha and Tanupriya Choudhury

Abstract In a developing country, the factors that improve the Gross Domestic Product (GDP) value should show an increasing trend in the graph. Agriculture is one such factor that contributes about 14% to the GDP of India. After the Green Revolution, India has taken a step ahead toward the growth of the agriculture industry. Growing population demands for a huge crop production. This research work focuses on the factor that accounts for crop growth. The crop data analysis is done for different seasons (i.e., Summer, Autumn, and Winter). The clustering technique of data mining is used to find interesting patterns to relate the area and production and how these factors have taken an upscale.

Keywords GDP · Agriculture · Clustering · Area · Production

1 Introduction

Agriculture is the foundation of the Indian economy. India has the second position worldwide after China in farm output. In the recent year, it has been noticed that the economic contribution of agriculture to India's GDP has declined. However, agriculture still remains the broadest sector of production. The land available for production is 159.7 million hectares. Wheat, rice, pulses, cotton, peanuts, fruits, and vegetables are the crops that make India among the top producers of the world. Being a developing country, India has the potential to enhance the production. There are several factors that affect the agricultural growth and the production.

D. Sharma · A. Sai Sabitha · T. Choudhury (✉)
Amity School of Engineering and Technology, Amity University Uttar Pradesh, Sector-125, Noida, Uttar Pradesh, India
e-mail: tchoudhury@amity.edu; tanupriya1986@gmail.com

D. Sharma
e-mail: dakshita.sharma96@gmail.com

A. Sai Sabitha
e-mail: saisabitha@gmail.com

© Springer Nature Singapore Pte Ltd. 2019

V. S. Rathore et al. (eds.), *Emerging Trends in Expert Applications and Security*,
Advances in Intelligent Systems and Computing 841,
https://doi.org/10.1007/978-981-13-2285-3_70

drksujatha23@gmail.com

Agricultural practices followed in our country are neither environmentally or economically sustainable. There is still a lack of good extension service like irrigation supply which is one of the main reasons for crop failure. The timely supply of inputs and transfer of outputs is also hindered, due to the poor network of roads. Crop yield varies significantly among Indian states. In comparison among the states, there is a difference between production grains per acre. The states larger in the area definitely account more to the production of crops. The distribution of food in India is very ineffectual as the transportation of agricultural products are highly affected, with inter-state and inter-district restrictions on marketing and movement of goods. Various studies state that the Indian agricultural policy should focus on improving rural infrastructure in the form of irrigation, flood control, transfer of knowledge on yielding techniques and developing more resistant seeds. Moreover, food packaging, storing and waste reduction are some of the factors that must be taken care of.

Low productivity is because of the various factors like size of land holdings is very small maybe because of land disputes, ceiling act, etc. The other factors are practice of old and traditional ways to bring the crop, inadequate knowledge upon the utilization of resources, increased cost of products the risk involved in whether the crop will do well or not, insufficient supply chains.

In 2012, the National Crime Records Bureau of India reported that there is an increase in farmer suicides which accounts for 11.2% of all suicides in India. Many reasons have counted for farmer suicides, such as monsoon failure, high debt burdens, genetically modified crops, government policies, public mental health, personal issues and family problems. This paper focuses on the issues behind the hindrance in crop production and to observe the different patterns. The main objective is to find the unique pattern in crop production in the years from 1996 to 2014 using data mining techniques.

2 Literature Review

Clustering is an unsupervised learning technique that finds a natural grouping of instances. This technique is performed to get a set of meaningful subclasses. The best quality of clustering is to find the hidden patterns. There are various clustering methods like partitioning, hierarchical, density based, grid based, model based, and constraint based.

K means clustering aims to partition "n" observations into "k" clusters in which each observation belongs to the nearest mean value cluster. It helps in finding clusters of comparable spatial extent. Clusters can have different shapes because of the expectation–maximization mechanism. There are different kinds of methodologies used by the researchers throughout the world in the field of agriculture. Some of the major researches are: Ramesh and Vishnu Vardhan analyzed the agriculture data from 1965 to 2009 in East Godavari district of Andhra Pradesh. They clustered the rainfall data into four clusters by adopting the K means

drksujatha23@gmail.com

Table 1 Data mining in crop production

S. no.	Author	Year	Application	Technique used
1.	R. Sujatha and Dr. P. Isakki	2016	Crop yield forecasting	Classification
2.	Niketa Gandhi and Leisa J Armstrong	2016	Impact of rainfall on rice crop yield	Association
3.	Niketa Gandhi et al.	2016	Predicting rice crop yield	Bayesian networks
4.	Niketa Gandhi et al.	2016	Rice crop yield prediction	Support vector machine
5.	D. Venkataraman	2016	Yarn price prediction	Advanced analytics
6.	Jharna Majumdar and Shilpa Anakalki	2017	Analysis of agriculture data	Multiple linear regression

clustering method. Multiple Linear Regression (MLR) is used to find the linear relationship between the dependent variable (rainfall) and the independent variables (year, area, production). The purpose of this study was to find suitable data models that may achieve high accuracy and generality in terms of yield prediction. Niketa Gandhi and Leisa J Armstrong used data from the government repository to study the use of Bayesian Networks to predict rice crop yield for Maharashtra. In this research work, 27 districts of Maharashtra were chosen. The parameters selected for the study were precipitation, temperature during the Kharif season, reference crop, production and yield for the years 1998–2002.

A comparative study was conducted between Naïve Bayes and Bayes Net classifiers to predict the accuracy, sensitivity, and specificity in crop yield production (Table 1).

3 Methodology

The methodology followed in this research is as follows:

(1) Data exploration
(2) Data cleaning
    (i) Removing missing values
    (ii) Outlier data
    (iii) Handling inconsistency
(3) Determining cluster parameters
(4) Cluster analysis.

drksujatha23@gmail.com

	A	B	C	D	E	F	G
1	State_Nam	District_Name	Year	Crop_Seas	Crop_Type	Area	Production
2	Assam	BARPETA	1997	Autumn	Rice	83560	1278
3	Assam	BONGAIGAON	1997	Autumn	Rice	29514	145.2
4	Assam	CACHAR	1997	Autumn	Rice	8278	10
5	Assam	DARRANG	1997	Autumn	Rice	68782	8582
6	Assam	DHEMAJI	1997	Autumn	Rice	12750	1.9

Fig. 1 Dataset used

3.1 Data Exploration

The data collected from a government site had 246,902 rows and 7 columns. The attributes are as follows.

- State name: the data is of the 29 states
- District name: almost all the districts of a particular state were covered.
- Year: the data is from 1996 to 2014
- Season: there were 5 seasons
- Crop type: crop produced in the particular region
- Area: unit in hectares
- Production: unit in tonnes (Fig. 1).

3.2 Data Cleaning

The task was to minimize the data so that clustering could give better results. Data cleaning was done in the following ways:

Removing missing values: In the dataset, there were many missing attributes which are not considered for analysis.

Outlier data: The data set consists of crop details of 29 states. The units of area and production had a wide range of values. The outlier data was removed and a particular range was selected for the following attributes.

Area-10,000–10,000 hectares

Production-100–10,000 tones

Handling inconsistency: Some data were inconsistent. The consistent data were selected for the seasons like autumn, summer, and winter. The data was further normalized using—Eq. (1). Both the parameters (area and production) were ranged from 0 to 10 as shown in Table 2.

Table 2 Normalized area and production

Value	Class (area)	Class (production)
0–4	Less land	Poor production
4–7	Average land	Average production
7–10	Huge land	Good production

drksujatha23@gmail.com

$$v' = \frac{v - \min_A}{\max_A - \min_A}(new\_\max_A - new\_\min_A) + new\_\min_A \quad (1)$$

V—value of area/production; Min—minimum value of the range; Max—maximum value of the range

New_max—maximum value of new range (10); New_min—minimum value of new range (0).

3.3 Determining Cluster Parameters

The clustering technique was used from Rapidminer. The value of 'k' was identified as '4' using Silhouette measure.

3.4 Cluster Analysis

Clusters were formed for different seasons. The scattered plots were plotted between the following:

(a) Cluster versus Production, (b) Cluster versus Area, (c) Cluster versus Year, and (d) Area versus Production. The following cluster models were obtained as

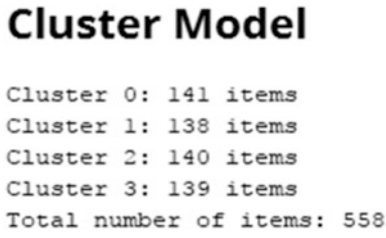


Fig. 2 Cluster model of autumn season

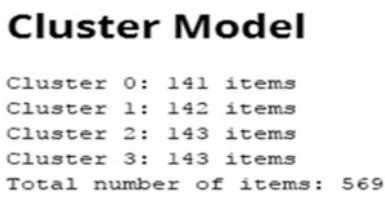


Fig. 3 Cluster model of summer season

drksujatha23@gmail.com

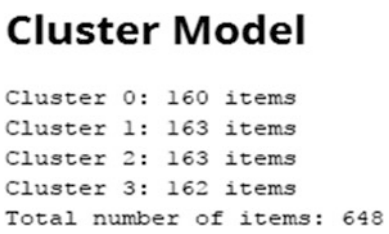


Fig. 4 Cluster model of winter season

shown below. The cluster model showed up an average object in each cluster. Each cluster shows certain states and crops. The value of the area and production is also similar in the clusters (Figs. 2, 3 and 4).

4 Result and Analysis

Season 1: Autumn

Figure 5 describes that rice was only produced in the state of Assam. In reference to the area available, the production in Bihar was poor. Paddy was produced in Bihar from the year 2011 to 2014 and the production was good.

Figure 6 shows that ragi was produced in Darjeeling district of West Bengal. Production of ragi was very high in 2009 but it suddenly declined in 2010. Maize was largely produced in that year. Hence, it may be the reason for less production of ragi. Paddy was produced in Orissa only in 1997 and the production was average.

Figure 7 narrates that in Bihar, the production of maize remained average from 1996 to 2014.The production of rice in few region of Bihar like Champaran, Munger had been always constant with respect to area. While in other districts, production was quite less as compared to the area.

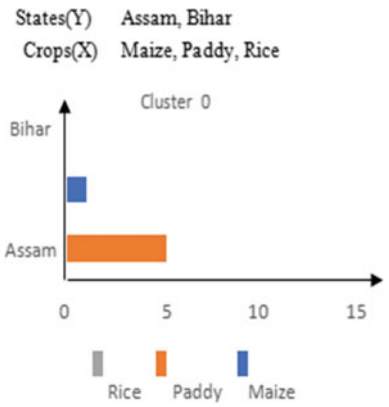


Fig. 5 Cluster 0 of autumn

drksujatha23@gmail.com

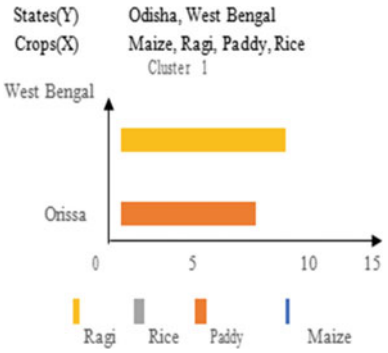


Fig. 6 Cluster 1 of autumn

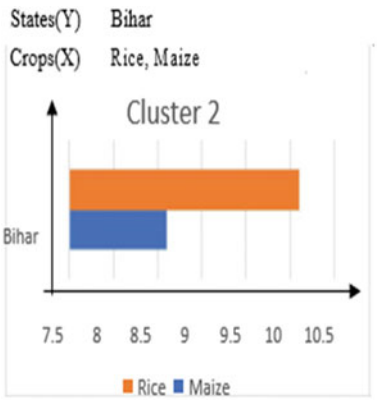


Fig. 7 Cluster 2 of autumn

Figure 8 portrays that the Bargarh region of Orissa produces groundnut. The production from the year 2004 to 2014 has remain average and the whole area is utilized for production. Ranchi region of Jharkhand produces ragi and the production has always been average since 2003.

Summary: The crops grown in autumn season are mainly Kharif crops (i.e., rice, paddy, maize, ragi, groundnut) and the states that are in constant production of these crops are Bihar, Assam, Jharkhand, Bihar, West Bengal, Orissa. We see that the production is not as much as the area available. Only in certain districts, the production has shown a high graph. The reason may be that in the autumn season there is not much supply of water. This is needed for the crops and hence, the production of Kharif crop is insufficient in many areas.

Season 2: Summer

Figure 9 depicts that in the Hooghly region of West Bengal, good production was observed from 2010 to 2014. Certain improvements were seen in the last decade. Paddy was grown in Assam as compared to the other states and the production was very good. Rice and sesame had average production in West Bengal. The crops have seen a rapid improvement in production.

drksujatha23@gmail.com

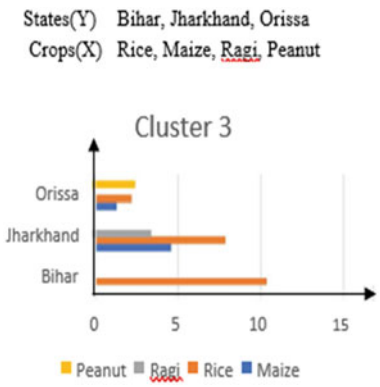


Fig. 8 Cluster 3 of autumn

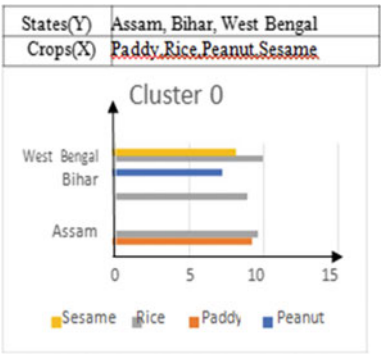


Fig. 9 Cluster 0 of summer

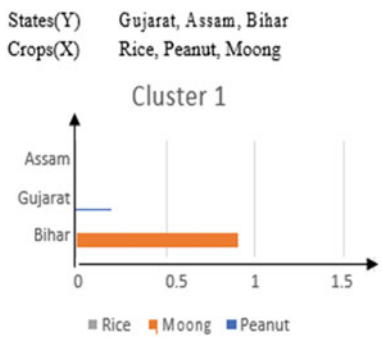


Fig. 10 Cluster 1 of summer

Figure 10 accounts that in certain districts of Gujarat, the area available for production of groundnut was quite less since 1997 and hence, the production was very low. The region of Assam had quite an average area and the production of rice was minimal. In Bihar, the production of moong was negligible in the year 2012–2014.

drksujatha23@gmail.com

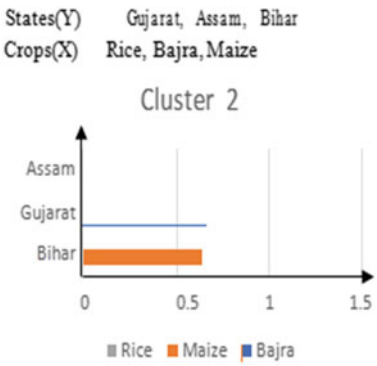


Fig. 11 Cluster 2 of summer

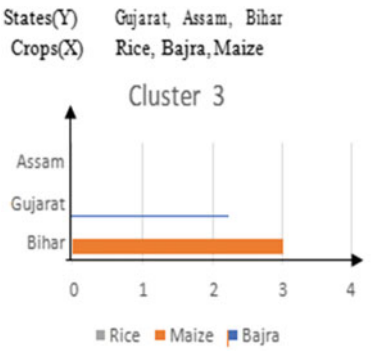


Fig. 12 Cluster 3 of summer

Figure 11 explains that in certain districts of Gujarat, the area available for production of bajra since the year 2000 was ample but the production was very low. The region of Assam had quite an average area and the production of rice was minimal. In Bihar, the production of maize had remained constant both according to area and production.

Figure 12 brings out that in Banas Kantha district of Gujarat, the area available for production of bajra was very huge but the production was minimal. The region of Assam had shown improvement in production of rice over the past few years. In Bihar, the production of maize was the maximal in the year 2013.

Summary: At the end of the summer season, the beginning of Kharif season starts and most of the Kharif crops (i.e. bajra, maize, rice, moong, paddy, sesame) are grown. So, we could see from the graphs that Assam, Bihar, and Gujarat had a significant production. Initially, the production was very low with respect to the area, but increased as the season progressed. In the starting clusters, it was seen that the ratio of area versus production was 10:2 but increased to 10:4 as the season ended.

drksujatha23@gmail.com

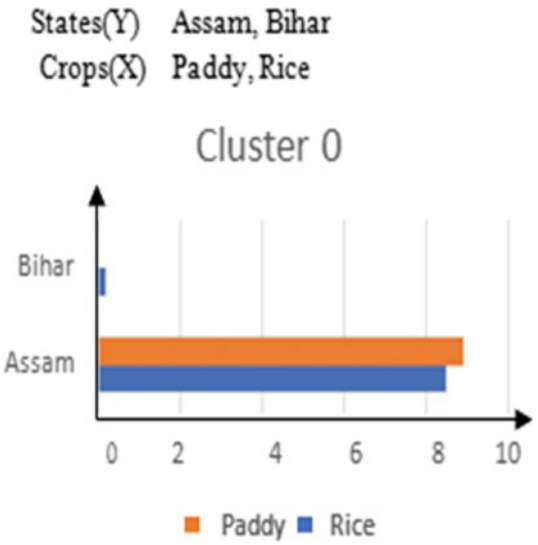


Fig. 13 Cluster 0 of winter

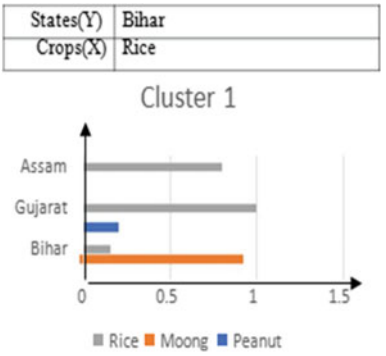


Fig. 14 Cluster 1 of winter

Season 3: Winter

Figure 13 describes that in Assam, from the year 2011–2014, the area available for production of rice was huge but the production was minimum. While in Bihar the production of paddy was minimal. In Assam, the production of rice was quite less as compared to Bihar.

Figure 14 depicts that the area available for production in Bihar was quite large, but the production was really negligible. It was the same scenario with all the districts. The regions which had the highest production also showed minimum production.

Figure 15 portrays that the production of Urad in Cuttack district of Orissa was very good according to the area. There were mostly utilized. Potato was produced in

drksujatha23@gmail.com

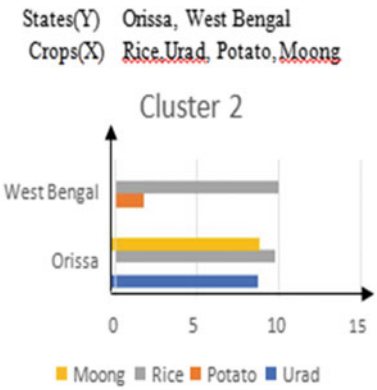


Fig. 15 Cluster 2 of winter

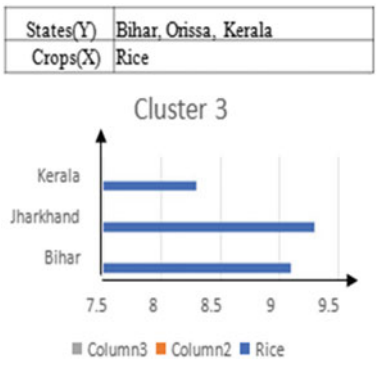


Fig. 16 Cluster 3 of winter

West Bengal and the production has been increasing since 1999. Ganjam district of Orissa had prominent production of moong crop.

Figure 16 narrates that Jharkhand had shown great production of rice over the years. The graph has been increasing since 1997. In Kerala and Bihar, the production was good. The area available for cultivation was enough to have good productivity.

Summary: In the winter season, there is the production of some Kharif crops (i.e., rice, paddy, moong, urad, potato). There was ample production of rice and paddy in the states of Bihar and Assam. While in other states production was low. It seems to be that the Kharif crops are not suitable for the winter climatic condition. The production of rice was good in Bihar, Jharkhand, and West Bengal as river Ganga flow through these regions or there is an ample supply of water.

drksujatha23@gmail.com

5 Conclusion

Clustering data mining technique was applied to the crop dataset. The various patterns were observed according to the seasons (autumn, summer, winter). The main findings were that in offseason for a crop the area was hugely available but the production was very low. It is because of improper irrigation facilities available. Kharif crops were grown in every season, but the production varied hugely. The regions which had good production of Kharif crop in offseason had a source of water through that area (Ganga flows through those regions).

In this research work, it was found out that the small-scale industries have increased rapidly hence, the farmer's involvement in farming activities has reduced. There is an increase in commercializing their product or crop. Thus, the production is very low.

The extended work can be carried for the Kharif and the rabi season. K-means clustering technique is used in this research work. Other techniques like hierarchical clustering and density-based clustering can be used to obtain the enhanced results.

drksujatha23@gmail.com

Predicting the Accuracy of Machine Learning Algorithms for Software Cost Estimation



Chetana Pareta, N. S. Yaadav, Ajay Kumar and Arvind Kumar Sharma

Abstract Today, the software cost estimation becomes rising region among many significant issues looked by programming improvement and programming industry. It is a necessary issue to foresee correct cost estimation keeping in the mind that the ultimate objective is to coordinate well spending arrangement. Usually, data mining is a mechanism towards analyzing information from exchange perspectives and compacting it into important information. Data can be utilized to grow pay, cut expenses or both. It empowers customers to separate information from several estimations or edges, characterize it and pack the associations recognized. This paper introduces a novel idea of building a model using ML algorithms into the existing software cost estimation models and simulated to predict the cost estimation parameters. The obtained model would be predicted the cost, effort, and schedule.

Keywords Software cost estimation · Machine learning algorithm Naive Bayes · WEKA

1 Introduction

Cost estimation is a method or a figure of the achievable cost of a thing, program or an undertaking, enrolled in the light of available information [1]. Cost estimation joins the strategy or procedures that help us in reckoning the honest to goodness and total cost that will be required for our item and is considered as one of the surprising and testing activities for the item associations. They will probably make programming which is unassuming and meanwhile pass on extraordinary quality [2].

C. Pareta (✉) · N. S. Yaadav · A. Kumar
JECRC University, Jaipur, India
e-mail: chinkyjaiswal0@gmail.com

A. K. Sharma
University of Kota, Kota, India
e-mail: drarvindkumarsharma@gmail.com

© Springer Nature Singapore Pte Ltd. 2019

V. S. Rathore et al. (eds.), *Emerging Trends in Expert Applications and Security*,
Advances in Intelligent Systems and Computing 841,
https://doi.org/10.1007/978-981-13-2285-3_71

drksujatha23@gmail.com

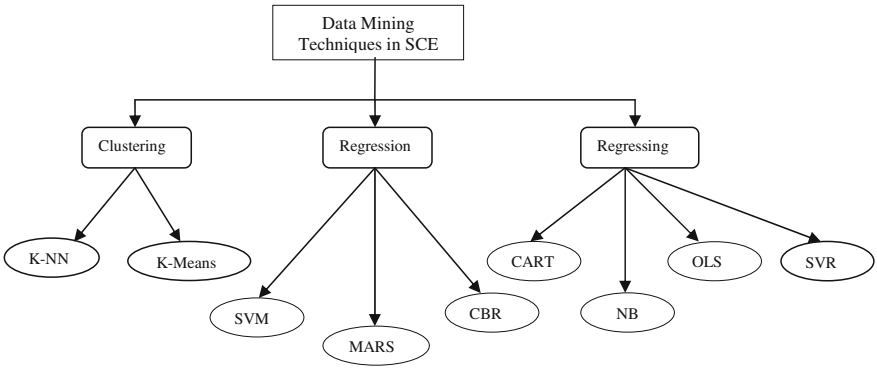


Fig. 1 Data mining techniques

Particular programming cost estimation models have been conveyed. Correct programming cost checks are essential to the two originators and customers. They can be utilized for making interest for proposals, contract exchanges, arranging, watching, and control. Cost estimation is a strategy or a figure of the conceivable cost of a thing, program or an undertaking, enrolled in view of open information. Correct cost estimation is fundamental for every sort of errand, on the off chance that we do not evaluate the undertaking really; the resulting cost of the undertaking is high all over it will be accomplished 150–200% more than the chief cost [3]. The item structures that we work with are intrinsically astounding and difficult to conceptualize. This multifaceted nature incites faults and distortions as a result increase the cost of programming. Programming estimations have for a long while been a standard apparatus for studying the nature of programming structures and the techniques that convey them. In any case, there are a few shortcomings utilizing the estimations as directors by and large rely upon estimations which they can without quite a bit of extending procure and work with. Noteworthy estimations are difficult to get and are inaccessible. The data information made in a programming system is tremendous and hard to work with. If suitable handling is done, it can be critical for various programming planning strategies and stages [4]. The software cost estimation has been improved with the help of different data mining techniques. Figure 1 shows some of the most popular data mining techniques in the context of software cost estimation.

2 Related Work

This area talks about a few research work, completed by many creators and masters in the field of programming cost estimation identified with information mining strategies. Some of them are as follows.

drksujatha23@gmail.com

Khan et al. [5] proposed a neural network technique for Software Effort Estimation which is investigated and reviewed in the perspective of MMRE and Predicate. They starting at now utilized Artificial Neural Network-based Effort Estimation strategies have displayed and evaluated. It was discovered that Functional Link ANN (FLANN) decreases the botch to the most outrageous purpose of control and decays the disperse quality than other ANN strategies. Sophatsathits et al. [6] showed different existing endeavor cost estimation procedures to analyze how the estimation should be possible in a more correct and productive way. This survey examined different existing programming's wander cost estimation methodology and estimations from the assignment life cycle position. Rani et al. [7] showed an overview of the three fundamental cost estimation strategies (work point, utilize case point, and line of code). These techniques show that how we could decide the cost of programming according to the essentials. For these circumstances, use case framework are useful when diverged from work point and loc. The line of code is not for the most part used. Likewise, use case point and limit point are extensively used. Mom et al. [8] showed a visual, versatile, and semantically solid intend to address fruitful and capable thing cost structures, visit outline assortments and business changes. Another component-based semantic model has been proposed for brought estimation reason. This model is based more than three submodels: feature-based association mapping, information mining and semantic showing. Gharehchopogh et al. [9]; showed an Artificial Bee Colony (ABC) calculation the reliance between components of COCOMO show has been assessed and the best a motivating force to assess the cost and effort of undertakings has been given. COCOMO-ABC exhibits were proposed to better measure and beat the COCOMO appear. Results exhibited that the COCOMO-ABC demonstrates has diminished the MMRE motivating force to around 1.77 times than the COCOMO show. Tannu et al. [3]; displayed a detailed survey of existing programming cost estimation models and systems. They were presented the favored angle and shortcoming of various cost estimation technique. This paper in the like way presents a piece of the material reasons that reason misguided estimation. They conveyed a vital and strong measure, we should upgrade our cognizance of programming wander attributes and their causal associations, make convincing strategies for assessing programming disperse quality and the cost estimation process ought to be totally engineered intentionally. Kumari et al. [10] displayed a similar division among the existing well-known models which was performed and the execution was examined and looked at in wording MMRE (Mean Magnitude of Relative Error) and PRED (Prediction). To create a significant and solid assessement, we should enhance our comprehension of programming venture properties and their causal connections, create successful methods for estimating programming unpredictability and the cost estimation process should be altogether organized and deliberately took after. Ziauddin et al. [11] displayed to utilize a method of reasoning model to upgrade the accuracy of programming effort estimation. Delicate strategy for thinking is utilized to fuzzily enter parameters of COCOMO II and the outcome is defuzzified to get the resultant effort. Triangular enrolment work has been used to create phonetic-specific regards. One can without quite a bit of

drksujatha23@gmail.com

extending develop a comparative model using trapezoidal or Gauss Bell support limits. Furthermore, there is still an edge of progress in the fuzzification process. Rathore et al. [12]; showed the chart of a couple of cost estimation models and strategies. To make a critical and reliable check, the cost estimation process ought to be totally planned and definitely many. It appears that all estimation strategies are particular for some particular sort of ventures. It would be best while evaluating the cost that more than one estimation technique be utilized to get a global see on the conceivable cost. Suresh et al. [13]; exhibited a flawless cost estimation technique recommended an approach to manage and improve the utility of the product cost evaluated by uncovering vulnerability (incomprehension of the venture and in addition in costing exactness) and decreasing the hazard that the gauge will be far not the same as the genuine cost. It can gives one kind of great confirmation. These things are effective. Dynamic programming cost estimation models are actualizing utilizing portfolio determination method. Sheta et al. [14] presented two new models for programming effort estimation utilizing fluffy rationale. This was a testing issue for programming venture administrator. They investigated the utilization of fluffy rationale as a delicate figuring method, which can improve the displaying procedure of the exertion. Two models motivated from the COCOMO and FP were created based on fluffy rationale. The created fluffy models actualized based the Takagi-Surgeon system. Pauline et al. [15] gave an issue the present strategy for evaluating limit centers that compel the powerful utilized of capacity indicates and proposed a change in the approach that should improve the precision. They displayed fluffy grouping strategies as a reason for building quality models. Observational approval for programming improvement exertion multipliers of COCOMO II is researched and the assessments for the cost drivers are portrayed.

3 Research Methodology

There are so many techniques available for software cost estimation but they are not very effective. Hence, there is more work is to be done using data mining and software engineering. It attempts to data predict good results to combined both of the field software engineering and data mining in the work. It generates the accurate cost of the project with the help of past dataset whose cost or effort is unknown and find out the common cot factors. The machine learning tool WEKA is used for this work. Figure 2 shows a functional diagram of the research methodology.

3.1 Dataset

In this work, a dataset has been used to simulate by using machine learning algorithms in WEKA.

drksujatha23@gmail.com

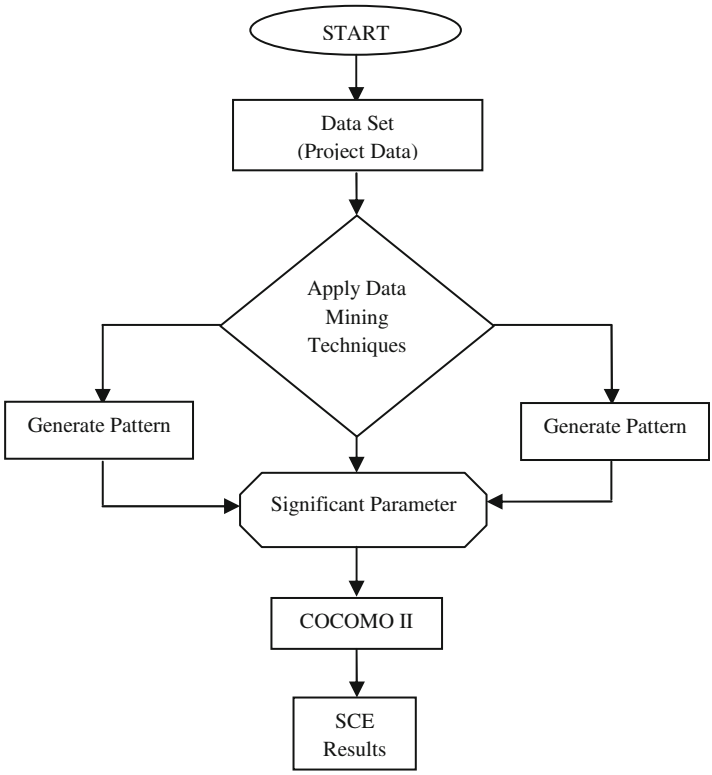


Fig. 2 Research methodology

3.2 WEKA Tool

WEKA remains for Waikato Environment for Knowledge Analysis. It is an exceptional machine learning instrument written in Java, made by Waikato University in New Zealand. The WEKA application permits novice customers an instrument to perceive covered information from database and record structures with simple to use options and visual interfaces. WEKA consolidates such countless learning calculations for data mining errands [16]. WEKA gives an instrument compartment of learning calculations and in addition to a structure inside which researchers could execute new counts without being stressed over the supporting system for data control and plan appraisal.

Nowadays, WEKA is seen as a state of the premium system in data mining and machine learning [17].

drksujatha23@gmail.com

3.3 Naive Baye's Algorithm

Naive Baye's is one of the popular machine learning algorithms. Naive Baye's classifier expects that the proximity of a particular component in a class is separated to the closeness of some other component. Naive Bayes is anything but difficult to assemble and especially helpful for vast informational indexes. Alongside effortlessness, Naive Bayes is known to outflank even exceedingly complex arrangement techniques. Naive Bayes classifiers are very versatile, requiring various parameters direct in the quantity of factors (highlights/indicators) in a learning issue. The Naive Bayes calculation is a straightforward probabilistic classifier that ascertains an arrangement of probabilities by checking the recurrence and mixes of qualities in a given informational index. The calculation utilizes Bayes hypothesis and accepts all credits to be free given the estimation of the class variable. This contingent autonomy suspicion once in a while remains constant in true applications, consequently the portrayal as Naive yet the calculation has a tendency to do well and learn quickly in different regulated arrangement issues. Naive Bayesian classifier depends on Bayes' hypothesis and the hypothesis of aggregate likelihood. For example, a characteristic item may be believed to be an apple in case it is red, round, and around 4″ in remove over.

Despite the fact that these highlights rely upon the presence of alternate highlights, a Naive Bayes classifier considers these properties to autonomously add to the likelihood that this organic product is an apple [18]. The favorable position of the Naive Bayes classifier is that it just requires a little measure of preparing information to gauge the methods and differences of the factors important for grouping. If free factors are unspecified, just the changes of the factors for each name should be resolved and not the whole covariance lattice. Rather than the Naive Bayes administrator, the Naive Bayes (Kernel) administrator can be connected on numerical qualities. The condition for Bayes hypothesis is given by

$$\mathrm{P(A/B)} = \frac{P\left(\frac{B}{A}\right)P(\Delta)}{P(B)}$$

In basic terms, a Naive Bayes classifier expects that the closeness (or nonappearance) of a particular component of a class is immaterial to the proximity (or nonattendance) of some other segment [19].

4 Experimental Evaluation

Here, in this work the machine learning algorithm, i.e., Naïve Bayes is used to compute the performance measures and the accuracy of software effort of proposed model and it is evaluated on the basis of MRE (Magnitude of Relative Error), RMSE (Root Mean Square Error), and MMRE (Mean Magnitude of Relative Error). Following Simulation Results of Naïve Bayes Algorithm Fig. 3 shown in below (Fig. 4).

drksujatha23@gmail.com

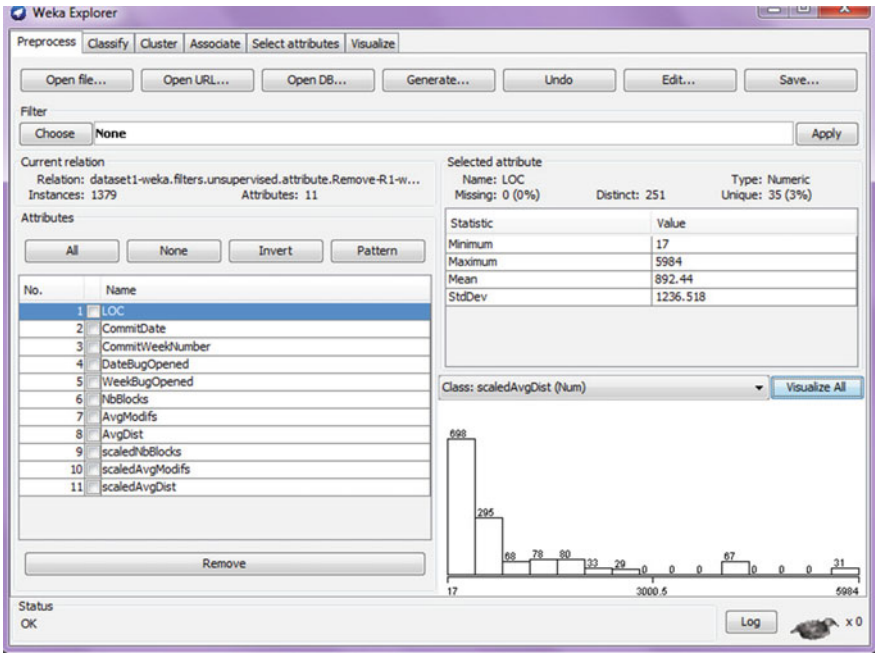


Fig. 3 Window shows all attributes of the dataset in WEKA

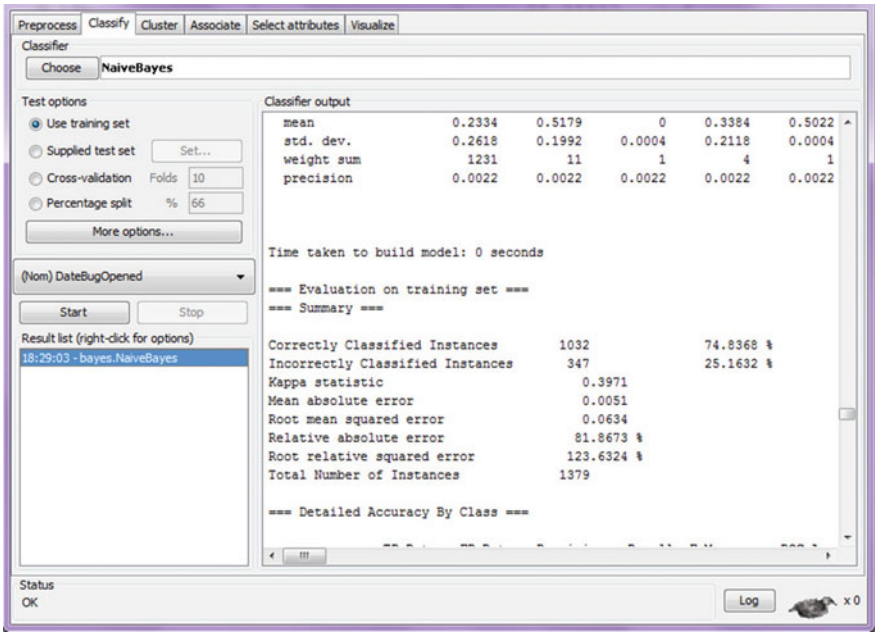


Fig. 4 Window shows simulation results of the Naïve Bayes algorithm

drksujatha23@gmail.com

5 Performance Evaluation

There are various systems for surveying the estimation exactness of programming effort proposed show, for example, Magnitude of relative (MRE) oversight [20]. In any case, decide the level of assessing blunder in an individual measure for each datum point as a venture. It is described as [10].

$$\text{MRE} = \frac{Predicted\ Value - Actual\ Value}{Actual\ Value}$$

RMSE—It stands for Root Mean Square Error. It is consistently utilized to measure of the difference between regards anticipated by a model or estimator and the qualities truly observed from the thing being exhibited or assessed. It is just the square establishment of the mean square bungle as seen in condition given below [10].

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Actual\ Value - Predicted\ Value)^2}$$

MMRE—It stands for Mean Magnitude of Relative Error. The relative error is given by the large mix-up in the observations apportioned by its genuine regard [21]. It is another measure and is the level of the total estimations of the relative errors, found the center estimation of over the N things in the "Test" set and can be created as [10].

$$MMRE = \frac{1}{n}\sum_{i=1}^{n}\frac{Predicted\ Value - Actual\ Value}{Actual\ Value}$$

PRED (N)—It is the third criteria utilized for the relationship and this reports the normal level of evaluations that were inside N% of the honest to goodness respects. It is typically utilized and is the level of conjectures that fall inside p% of the genuine, suggested as PRED (p), k is the measure of activities where MRE is not precisely or identical to p and n is the measure of attempts [10].

$$\text{PRED}\ (\text{p}) = \text{k}/\text{n}$$

The performance evaluation of the Naïve Bayes algorithm for the corresponding dataset utilizing WEKA tool is presented in Table 1 (Fig. 5).

The accuracy of machine learning classification of Naïve Bayes algorithm for software cost estimation on the given dataset is presented in Table 2.

The performance evaluation metrics of the Naïve Bayes algorithm is shown in Fig. 6.

It shows correctly classified and incorrectly classified instance by using Naïve Bayes. The accuracy of this model is achieved 97.2342%.

drksujatha23@gmail.com

Table 1 Result analysis of Naïve Bayes algorithm

Correctly classified instances	1236	89.6302%
Incorrectly classified instances	143	10.3698%
Total number of instances	1379	

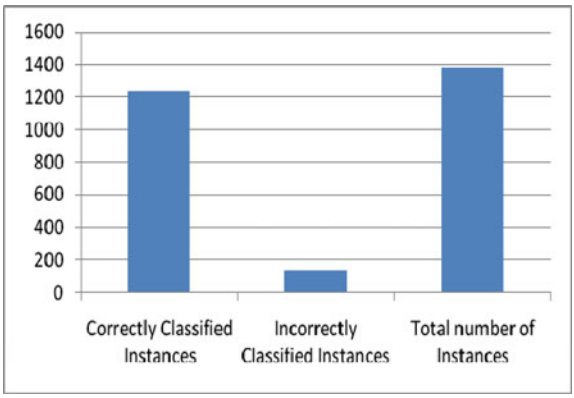


Fig. 5 Performance analysis of Naïve Bayes algorithm

Table 2 Accuracy of Naïve Bayes algorithm

Kappa Statistic (KS)	0.4943
Mean Absolute Error (MAE)	0.0027
Root Mean Squared Error (RMSE)	0.0499
Relative Absolute (RA) (%)	42.9605
Root Relative Squared Error (RRSE) (%)	97.2342

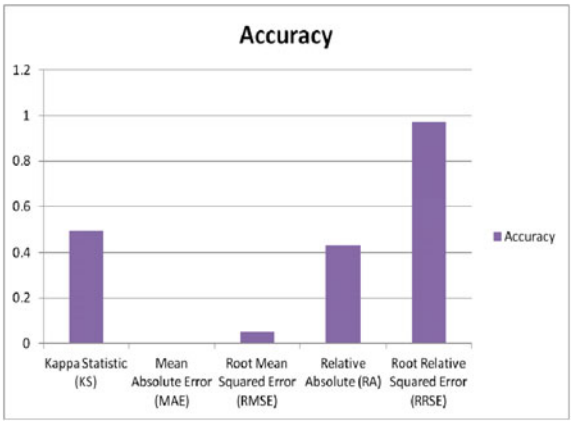


Fig. 6 Performance evaluation metrics of Naïve Bayes

drksujatha23@gmail.com

6 Conclusion

This paper focuses on the current programming estimation strategies. In like manner, displayed establishment data on programming venture models and programming estimations to be used for effort and cost estimation. In this paper, the gathering figuring's for data mining are used as a piece of this work by using the Naive Bayes table. The calculations have been connected for WEKA device and demonstrate the great execution and great exactness here. The Naive Bayes gives the best number of occasions 1379. In future, assorted computations can be associated with particular unmistakable data to predict the results in programming cost estimation.

References

1. Sharma N et al (2012) Incorporating data mining techniques on software cost estimation: validation and improvement. IJETAE 2(3):301–309
2. Shekhar S et al (2016) Review of various software cost estimation technique. IJCA 141 (11):31–34
3. Tannu et al (2014) Comparative analysis of different software cost estimation methods. IJCSMC 3(6):547–557
4. Pon Periasamy AR et al (2017) Application of data mining techniques in software engineering. IJARCSSE 7(3):304–307
5. Khan MW et al (2014) Neural network based software effort estimation: a survey. IJANA 5 (4):1990–1995
6. Sophatsathit P et al (2014) An exploratory survey of phase-wise project cost estimation techniques. AUJT 18(1):36–47
7. Rani J et al (2014) Comparison of cost estimation technique. IJARCSSE 4(5):1005–1009
8. Ma YS, Sajadfar N et al (2014) A feature-based semantic model for automatic product cost estimation. IACSIT 6(2):109–114
9. Gharehchopogh FS et al (2014) Artificial bee colony based constructive cost model for software cost estimation. JSRD 1(2):44–51
10. Kumari S et al (2013) Performance analysis of the software cost estimation methods: a review. IJARCSSE 3(7):229–238
11. Ziauddin et al (2013) A fuzzy logic based software cost estimation model. IJSEA 7(2):7–18
12. Rathore S et al (2013) Review on cost estimation methods for software development. IJR 2 (3):132–134
13. Suresh N et al (2013) Intelligent models creation for cost estimation using clustering techniques. IJETT 4(6):2350–2355
14. Sheta AF et al (2013) Software effort estimation inspired by COCOMO and FP models: a fuzzy logic approach. IJACSA 4(11):192–197
15. Pauline M et al (2013) Comparison of available methods to estimate effort, performance and cost with the proposed method. IJEI 2(9):55–68
16. Kaur K et al (2016) Review of data mining with Weka tool. IJCSE 4(8):41–44
17. Sharma TC et al (2013) WEKA approach for comparative study of classification algorithm. IJARCCE 2(4):1925–1931
18. Patil TR et al (2013) Performance analysis of Naïve Bayes and J48 classification algorithm for data classification. IJCSA 6(2):256–261

drksujatha23@gmail.com

19. Vijayarani S et al (2015) Linear disease prediction using SVM and Naïve Bayes algorithm. IJSETR 4(4):816–820
20. Tayraj V et al (2017) Exploration of effort estimation techniques. IJIACS 6(5):98–103
21. Nagpal G et al (2012) A comparative study of estimation by analogy using data mining techniques. JIPS 8(4):621–652

drksujatha23@gmail.com

Cloud Computing Research Issues, Challenges, and Future Directions



Dhirender Singh, R. K. Banyal and Arvind Sharma

Abstract Cloud computing could be a hot research area among the researchers in today's world. Cloud computing is thought on be a promising resolution for mobile computing as a result of many reasons for quality, portability, and communication. The need for mobility in cloud computing has given the worth to mobile cloud computing. The cloud computing during this paper has been explored the variety of mechanism for providing data security in order that cloud computing would be widely accepted by the variety of many users. Also, this paper presents an overview of cloud computing research issues, challenges, and future directions.

1 Introduction

Cloud computing is a web-based service. It is a strategy for giving on-demand services, expandable and perfectly elastic software services using the internet. Cloud computing is scalable and unambiguous like we can increase and decrease the resources anytime. Cloud computing could be a consistent access to nearly limitless resources. Cloud computing makes a good vary of solutions obtainable to us. A cloud could be a style of parallel and distributed system. It comprises the gathering of interconnected systems. It is virtualized computers that are dynamically provided together or additional computing resources. Cloud computing could be a new manner of considering IT services and how resources may be utilized in a broad way. Solutions and services that are bought and consumed in real time over the online are cloud services. For example, once we store our pictures online, we have a tendency to use webmail or social networking website; we have a tendency to be using a cloud computing service. Cloud computing may well be a distributed

D. Singh · R. K. Banyal
Rajasthan Technical University, Kota, India
e-mail: dhirendersingh92@gmail.com

A. Sharma (✉)
University of Kota, Kota, India
e-mail: drarvindkumarsharma@gmail.com

© Springer Nature Singapore Pte Ltd. 2019

V. S. Rathore et al. (eds.), *Emerging Trends in Expert Applications and Security*,
Advances in Intelligent Systems and Computing 841,
https://doi.org/10.1007/978-981-13-2285-3_72

drksujatha23@gmail.com

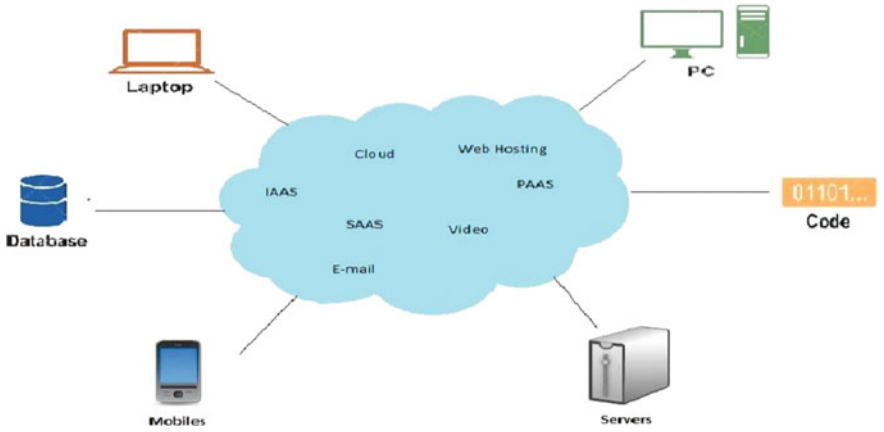


Fig. 1 Cloud computing

style that centralizes server resources on a climbable platform thus on provide on-demand computing resources and services due to the unexampled success of internet in previous couple of years, computing resources is presently plenty of ubiquitously available. Cloud computing has attracted the enormous companies like Google, Microsoft, and Amazon and thought of as a good influence in today's data technology business (Fig. 1).

The rest of the paper is organized into different sections as follows: Sect. 2 presents a brief background of cloud computing and its services. Section 3 discusses a brief summary of literature review. Section 4 presents research issues, challenges, and future directions. The conclusion is shown in Sect. 5 while references are mentioned at the last.

2 Cloud Computing—A Background

Cloud Computing was coined by the Google CEO Eric Schmidt. Cloud computing services are usually divided into three categories (Fig. 2).

- **Software as a service (SaaS)**: SaaS provides third-party software to the clients as a service on-demand. Clients do not get to install any software on its aspect. It will use the software provided by the third party. The client uses the applications anytime and anyplace.
- **Platform as a service (PaaS)**: It provides the environment and the different types of tools for creating online applications.
- **Infrastructure as a service (IaaS)**: It is a diversity of cloud computing that has virtualized computing resources over the internet. It permits existing applications which run on cloud supplier's hardware.

drksujatha23@gmail.com

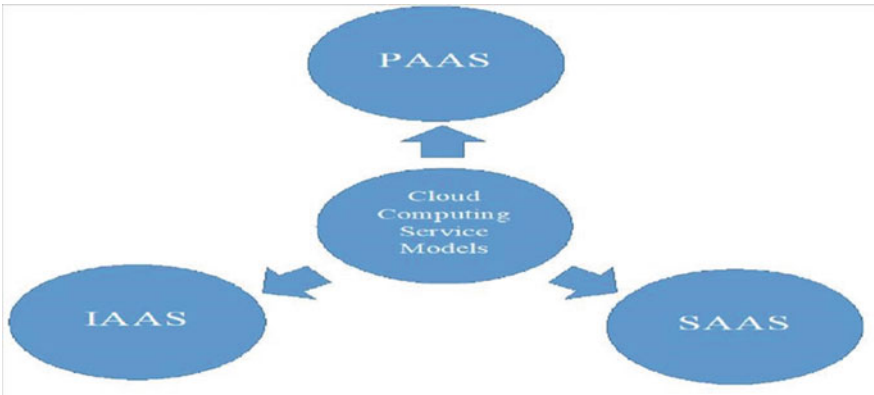


Fig. 2 Service models

Cloud Computing is divided into different categories. The categories are as under (Fig. 3):

- **Public Cloud**: Public cloud is a cloud which is publically available to all. It may be free or pay per use.
- **Private Cloud**: Private cloud which is dedicated to the single organization.
- **Hybrid Cloud**: Hybrid cloud is the combination of two or more cloud.

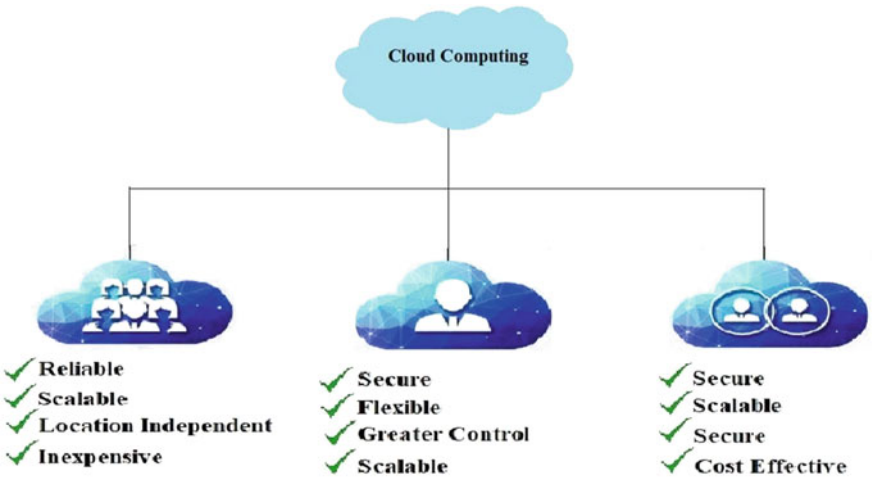


Fig. 3 Types of clouds

drksujatha23@gmail.com

3 Related Work

Several works have been carried out by many researchers in the domain of cloud computing on the basis of past research papers and literatures. Within the last past years, cloud computing has developed from being a promising business plan to at least one of the fastest developing fragments of the IT business.

They found some prime security attacks and supply attainable solutions for clouds: XML signature wrapping attacks, browser security, and vendor lock-in. They need been known problems and evaluated existing attainable solutions to those problems to settle on the most effective one among them mentioned in their paper [3]. A model for utility composition in cloud computing, this version was based mostly altogether on an illustration of manipulating ideas of an application domain via the bpm approach, particularly operations and static principles wont to describe the properties of business processes to assess the performance of this idea, a sequence of experimental tests became dispensed. When compared with a not composite model (NC) because of the usual approach on this place, the results showed that AC surpasses NC in terms of response time to time with the smallest complexity [9, 10]. A framework to unravel numerous moral and security aspects associated with cloud computing [11]. A Hadoop cluster uses the MapReduce framework to style the encryption service system. In addition, they offer a whole encryption scheme, a selective encryption scheme, and a partial encryption scheme to inscribe the video data. As a result, it not solely improves the speed of video encryption but additionally optimizes the video encryption strategy. Moreover, the user will choose the encryption scheme consistent with their necessities. The experimental results show that the video encryption service that they offer will meet the wants of high-speed, security, and so on [8]. Security risks associated issues in cloud computing and broadminded steps that an enterprise will like scale back security risks and defend their resources. They have also clarified cloud computing qualities/advantages, shortcomings, and appropriate regions in data chance management. In expansion, it cowls the advantages and disadvantages of the methodology of cloud computing. Additionally, it tackles the necessary side of security involved challenges that the researchers and authors face among the protection of cloud computing [6]. The issues faced by online education interaction and constructs a web education interactive platform model. By analyzing the platform functions and data processing, the paper realizes the high strength of online instructors and learner's interactive functions [5]. Most of the research work offers together with the identification of challenges within cloud forensics and between addition; the projected choices in retentive at intervals composition rely upon Cloud Service Supplier (CSP) for rhetorical inquisitors. The dependence on CSP includes the gathering of data for the forensics methodology and there is conjointly a chance for sterilization data that affects the complete investigation methodology. For mitigating the dependency on CSP, a brand new model for grouping forensic evidence outside the cloud atmosphere is developed [1]. The comprehensive review of this progressive and thus the newest developments on mobile cloud computing

drksujatha23@gmail.com

underneath the 5G era, which helps the beginning stage researchers to possess an outline of the present solutions, techniques, and applications, investigate the open analysis problems and future challenges in this domain [9]. A cloud-style reference model contains an enormous variety of security controls and best practices, and a cloud security assessment Model-Cloud-Trust estimates high-level safety metrics in accordance with quantifying the degree concerning confidentiality then integrity offered by using a CCS or cloud service provider (CSP). Cloud-Trust is used in imitation of assessing the protection level regarding four multi-tenant IaaS cloud architectures equipped with diverse cloud safety controls. Results show the probability of CCS penetrate on is high if the smallest amount set of security controls are implemented. CCS penetration likelihood drops considerably if a cloud defense thorough security style is adopted that protects virtual machine (VM) photos at rest, strengthens CSP and cloud tenant supervisor access controls that employ altogether totally different network security controls to scale back cloud network surveillance and discovery of live VMs [4]. A generic framework to handle such security at the first stages of the CSDLC. This framework aims at adding an additional security level at the first stages of the CSDLC that has been more illustrated by a case study showing the relevancy of the framework [2]. Investigated existing answer for SaaS model of cloud computing and explore the varied flaw in context of security. Here, the work concludes with the comparative study of the various existing answers and address the common issues and excuses [7].

4 Research Issues, Challenges, and Future Directions

Cloud is a set of technology, process, individuals, and business development. Like all other technology, individuals, business development, and cloud too have vulnerabilities. The accompanying are different types of the vulnerabilities in a cloud. Some of the open issues and threats that need critical consideration are as per the following:

- **Data Breach**: Confidential, sensitive, and protected data is obtaining by the unwanted user or a hacker. For example, WWE wrestling was recently the victim of a security breach that leaked the private information to three million users.
- **Denial of Service (DoS)**: Denial of service attack is to shut down the systems or network in order to overpower the victim resources and make it difficult for the authorized user to use.
- **Internet Protocol**: Many susceptibilities inside in IP such as ARP spoofing, IP spoofing, and DNS Poisoning are real threats.
- **Malicious Insider**: A decided insider can discover more approaches to attack and cover the track in a cloud situation.

drksujatha23@gmail.com

- **SQL Injection Vulnerabilities**: Vulnerabilities, as an example, OS injection, SQL injection, and LDAP injection at the management layer will cause important problems for completely different cloud customers.
- **API and Browser Vulnerabilities**: Any weakness in cloud supplier's API or Interface represents a major risk once combined with social engineering or browser-based attacks; the harm is often vast.
- **Changes to Business Model**: Cloud computing can be a critical change to a cloud consumer's business model. IT department and business need to adjust or confront introduction to chance.

There are some ways through which we can tackle these security issues.

- **Scanning for Malicious Activities**: End-to-end encryption while exceedingly prescribed prompts new dangers, as encrypted data cannot be read by the Firewall or IDS. In this way, it is imperative to have suitable controls and countermeasures to moderate dangers from malevolent programming going through encryption.
- **End-to-End Encryption**: The data in a cloud delivery model might navigate through various geographical areas; it is basic to encrypt the information end-to-end.
- **Validation of Cloud Consumer**: The cloud supplier needs to avoid potential risk to screen the cloud consumer to prevent important features of cloud being utilized for harmful attack purposes.

5 Cloud Computing Application Areas

There are numerous applications of cloud computing. The applications are:

- **Educational Institutions**: Cloud computing has definitely revolutionized the educational sector. The modern face-to-face classroom methods are added and added being replaced with alternative cloud-mediated exercises like smart classes using aesthetic and auditory illustrations.
- **Cloud Computing in Business**: The business delivery model demonstrate gives a client encounter by which software, hardware, and network resources are ideally utilized to give creative administrations over the Web, and servers are provisioned as per the sensible needs of the administration utilizing progressed automated tools. For associations right now utilizing traditional infrastructures, a cloud will empower clients to expend IT resources in the server farm in ways that were never accessible.
- **Cloud Computing in Medical Fields**: Within a hospital, so inside the bulk of medical practices, patient charts, and medical histories are usually unbroken within a system. Cloud computing will facilitate easier access and distribution of knowledge among the various medical professionals who would possibly are available involved with every individual patient.

drksujatha23@gmail.com

6 Conclusion

In this paper, we discussed the overview of cloud computing. Cloud computing is rising technology that is growing very speedily. Cloud computing provides smart measurability, service on-demand, computing capability, and so on. However, it additionally brings different types of problems at security, legal problems and lots of additional. Finally, this paper focuses on security, resources, and so on.

References

1. Alex ME, Kishore R (2017) Forensics framework for cloud computing. Comput Electr Eng 60:193–205. https://doi.org/10.1016/j.compeleceng.2017.02.006
2. Aljawarneh SA, Alawneh A, Jaradat R (2017) Cloud security engineering: early stages of SDLC. Futur Gener Comput Syst 74:385–392. https://doi.org/10.1016/j.future.2016.10.005
3. Alshammari A, Alhaidari S, Alharbi A, Zohdy M (2017) Security threats and challenges in cloud computing. In: 2017 IEEE 4th international conference on cyber security and cloud computing, pp 46–51. https://doi.org/10.1109/cscloud.2017.59
4. Gonzales D, Kaplan JM, Saltzman E et al (2017) Cloud-trust-a security assessment model for infrastructure as a service (IaaS) clouds. IEEE Trans Cloud Comput 5:523–536. https://doi.org/10.1109/TCC.2015.2415794
5. Jiugen Y (2017) Research on interactive application of online education based on cloud computing and large data, pp 593–596
6. Kajaree D, Behera R (2017) A survey on web crawler approaches. Int J Innov Res Comput Commun Eng 5:1302–1309. https://doi.org/10.15680/IJIRCCE.2017
7. Malgey S, Chauhan P (2016) A review on security issues and their impact on cloud computing environment. Int J Adv Res Comput Commun Eng 5:249–253. https://doi.org/10.17148/IJARCCE.2016.5653
8. Pei D, Guo X, Zhang J (2017) A video encryption service based on cloud computing
9. Sadok L, Okba K, Oueslati W (2017) Management by composition of applications, pp 144–150
10. Skourletopoulos G, Mavromoustakis CX, Mastorakis G et al (2017) Advances in mobile cloud computing and big data in the 5G era, p 22. https://doi.org/10.1007/978-3-319-45145-9
11. Surbiryala J, Li C, Rong C (2017) Framework for improving security in cloud computing. IEEE 260–264

drksujatha23@gmail.com

Social Big Data Analysis—Techniques, Issues and Future Research Perspective



Pranjali Borgaonkar, Harish Sharma, Nirmala Sharma and Arvind Kumar Sharma

Abstract Big Data increases such heaps of consideration in about each realm. All major square measures are serving with massive information because the immense quantity of data is being generated and utilized by the user solely on every day. Social media is one among all the cases wherever massive information plays a vigorous role and wishes to be handled in a correct means. Daily numerous reasonably information in kind of multimedia system is increasing day by day. In this paper, a study is presented on Big Data and its perspectives on social media. Moreover, this paper deals with numerous essential techniques evolved in past literatures. Also, some major problems and challenges arise in Big Data are discussed.

1 Introduction

Big Data became an awfully common analysis topic in numerous areas. One among them is social media. Everyday information is increasing exponentially from all its major sources like sensors, videos, pictures, social media posts, etc. Dataset generated by distinct sources are not straightforward to store, manage and analyse with ancient ways. Today, social websites area unit enjoying a major role in generating the explosive growth of data. With the advancements in technology, most are being a region of social network. Each web user expends around 3–5 h daily on totally different social media. Facebook precipitates four million posts every minute whereas YouTube originates three hundred hours of videos every minute (Fig. 1).

P. Borgaonkar · H. Sharma · N. Sharma
Rajasthan Technical University, Kota, India
e-mail: pranjal.cse.2892@gmail.com

A. K. Sharma (✉)
University of Kota, Kota, India
e-mail: drarvindkumarsharma@gmail.com

© Springer Nature Singapore Pte Ltd. 2019

V. S. Rathore et al. (eds.), *Emerging Trends in Expert Applications and Security*,
Advances in Intelligent Systems and Computing 841,
https://doi.org/10.1007/978-981-13-2285-3_73

drksujatha23@gmail.com

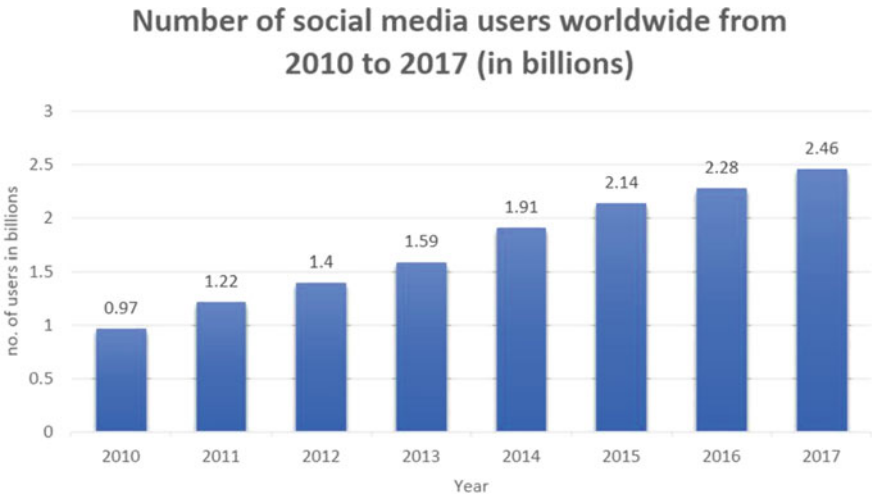


Fig. 1 Number of worldwide social media users (2010–2017)

Twitter conjointly sends five hundred million tweets on a daily basis. A study says around 7910 Exabytes (EB) of information was produced until 2015. Data made by social media could exceed by forty Zettabytes (ZB) by 2020 (wersm, n.d.).

2 Big Data—A Background

Big Data term describes the information generated therefore exemplified technologies like sensors, communications and computation. Big Data may add up as good as by seven V's. These V's are Volume, Veracity, Variability, Visualization, Variety, Velocity and Value.

2.1 Big Data Technologies

Hadoop

Hadoop is programming structure developed to support distributed computing surroundings to process large data blocks. Originally, Hadoop was a part of Apache Nutch project and then partitioned off into Hadoop subproject. Hadoop is a code framework where an application breaks down into numerous components and distributes them across the nodes in the distributed system. The present Apache Hadoop framework contains the Hadoop Kernel and varied modules like Hadoop Common, HDFS, Hadoop Yarn and Hadoop MapReduce. Hadoop ecosystem

drksujatha23@gmail.com

comes along with extra software packages like Apache Hive, Apache Pig, Apache Spark, Apache flume, Apache Sqoop, etc.

HDFS and MapReduce are described as follows:

HDFS

Hadoop concentrates on storage system known as Hadoop Distributed File System aka HDFS with fault tolerant capabilities. HDFS will store massive amounts of knowledge, rescale incrementally and survive the failure of significant elements of the storage infrastructure while not losing data.

HDFS architecture comprises Master–Slave architecture with the following entities:

- **Namenode**: The framework that contains namenode works as 'Master Node' and performs the distinctive task as dealing with the namespace gives the customer access to various documents, different operations sort of open, close, renaming registries and so forth. By and large, namenode stores metadata of HDFS, i.e. index tree of record framework, and tracks the document over the groups. Namenode is arranged with huge random-access memories (RAMs).
- **Datanode**: This is actual element also known as 'Slave Node' which is responsible for storing original data in HDFS. Namenode and data node coordinate by keep communicating with each other. Major operations of datanode are block creation, deletion and replication based on namenode instruction. Datanode is configured with High Hard Disks (HDDs).
- **Blocks**: Information that stored into the record in HDFS are secluded into a few fragments known as 'Blocks' having 64 bits as default measure which can be changed as need to be.

Different elements are otherwise called HDFS daemons are node administrator, resource manager, secondary namenode, yarn daemon, etc. Primary targets that HDFS serves are generally for management of large datasets, reduction of network traffic and throughput increment and finally fault tolerant and recovery behaviour.

MapReduce

MapReduce is a programming framework that provides processing techniques. The framework allows the jobs to split the input dataset into small nuggets and let run these parallelly. Two main functions of Hadoop MapReduce are as follows:

- **Map**: Map or mapper function takes the input datasets and converts them into intermediate dataset of key-value pair (Table 1).

Table 1 MapReduce input output

	Map	Reduce
Input	(k1, v1)	(k1, list (v1))
Output	list (k2, v2)	list (k3, v3)

drksujatha23@gmail.com

- **Reduce**: Reduce or reducer function takes input from mappers as input and reduces them by integrating data-tuples into an inadequate set of tuples. Reduce task usually comprises two task: 'shuffle' and 'reduce'.

2.2 Big Data Applications

Enormous data applications in medicinal services are the most well known now daily. As the information developing day by day, it is troublesome to assess size and improvement rates, by conventional measures. Estimated size of clinical information remains around 150 Exabytes till 2011, with a development rate in the vicinity of 1.2 and 2.4 Exabytes for each year. Transportation industry brings Big Data into service to help government, private and individual transports to serve the significant advantages and help them to make information-based and quick decisions. Private transports use Big Data mining to comprehend suburbanites and to mechanical upgrades and revenue management whereas government utilizes Big Data in traffic control, route planning and to enforce sharp and quick transportation systems. Some key aspects of Big Data in public sector are improving transparency between government officials and citizens with good decision capabilities and cost reduction [3]. For instance, different government tax firms make utilization of Big Data techniques to store personal data of customer which is replicated all over public sectors thus reducing errors by already filled application forms and this speed up process time [9]. The Internet itself fills in as the framework of a semantic web. Since the gigantic part of data is still now unstructured, there is a huge inspiration for arranging it into organized structures to derive related ideas and relations, with a specific end goal to computerize thinking.

3 Related Work

In this section, we discuss the various works have been carried out by the many researchers in the domain of Big Data and social media, which are as under.

Hameed et al. [4] presented a scientific classification of social sites assaults and gave our writing overview comes about that assistance to sort conceivable assaults and favoured contraventions to barrier against these assaults. Until the point when additionally see, this arrangement depends on the sort of correspondence utilized and aggressor' as objectives. Further, writers were zealous on exploring the security of social destinations for clients. In future, they intend to direct an appraisal on various informal community locales clients at various association levels to decide the social destinations assault on daily life. Liu et al. [7] presented three diverse positioning frameworks in China on WeChat Public accounts and break down how they assess and evaluate Big Data produced however web-oriented social

drksujatha23@gmail.com

networking with their unmistakable highlights: extensive, consolation, local. Facilitate recommendations are proposed for the positioning frameworks with enormous information created from online social networking stages also. In this aggregate research, three WeChat open record positioning frameworks are circumspectly contemplated and dissected. Be that as it may, it doesn't deplete all the positioning frameworks because of the obvious infeasibility; neither does it allude to other option computerized web-based social networking stages. Authors trust that exploration paper could be additionally advanced and enhanced later on when a portion of the information flaws they had delineated beforehand are rejected. Lee et al. [6] contrasted with past work, our work speaks to a progression in precision of appraisals, expectation of future influenza action precisely and a capacity to consolidate huge social information and watched CDC information to construct prescient models. The proposed model can foresee present and future flu exercises with high precision 2–3 weeks speedier than the conventional influenza observation framework can. For future considerations, they might be showing research how specifies of various side effects is connected with the real influenza levels and whether it could be utilized to enhance the flu movement estimate. Additionally, they should enhance influenza gauge precision by grouping tweets into numerous classifications (well-being, news, advertisements and so on.) and by applying fluctuating weights on various sorts of tweets because the quantity of posts discussing one influenza occurrence can shift contingent upon the classification of the tweet. Cheung and She [2] examined 3,152,344 pictures by 7,450 clients from Fotolog and Flickr, two picture situated interpersonal organizations. It is watched that clients who share outwardly comparative pictures will probably have a similar sexual orientation. A sight and sound huge information framework that uses this wonder is proposed for client sexual orientation recognizable proof with 79% precision. These discoveries are helpful for data or administrations in any informal organization with serious picture sharing. Yadav et al. [14] recommended the consumption of QR code idea. Promote in this survey they talked about the approach which is utilized for the security of individual information over web social networking and furthermore keep from the phony ids. In future, the proposed approach will give the best outcomes in picture encryption, and every single social medium scrambles the information utilizing QR code plot. Clients only need to exchange their QR codes and then the pictures are downloaded by any client. Nambisan et al. [8] described the undiscovered and untreated depressive issue that have turned into a genuine general medical problem, and this is pervasive among individuals of any age, sexual orientation and race. Web-based social networking destinations, for example, Twitter, have turned into a noteworthy scene for individuals to express/unveil their considerations and sentiments. The tweets from these miniaturized scale blogging destinations could be utilized to screen for and conceivably distinguish sadness. In this investigation, they constructed their examination discussing research on depressive issue, which shows the basic hugeness of tedious musings and ruminating conduct of individuals with misery. Immonen et al. [5] proposed arrangement enhances business basic leadership by giving continuous, approved information for the client. The arrangement is approved with a

drksujatha23@gmail.com

mechanical case in which client knowledge is separated from web-based social media information so as to decide the consumer loyalty concentrating on the nature of an item. Subasinghe et al. [11] proposed an alternate idea on chance recognizable proof of online networking systems utilizing Hidden Markov Models (HMMs) based on behavioural usage patterns. Unauthorized users are identified based on characteristic activity sequences. Here, a client behavioural model is worked for the online social media system customers utilizing their point by point action logs involving use designs over years. The neuro-fuzzy model gave an 84% precision to recognize true legitimate clients while giving 62.5% exactness to distinguish doomed clients. This gives a general exactness of recognizing substantial clients (honest to goodness or ill-conceived) of 76.85% and a false positive rate of 10.65% and a false negative rate of 12.5%. Future research work would be extended to test the model with more social media consumer data relating to different applications applying to both assessing and training phases carrying out further experiments to optimize our model. Smith et al. [10] taken a gander, as of now, issue from another point of view, in what manner can the client pick up attention to the by and by applicable part Big Data that is freely accessible in the social web. The measure of customer-created media exchanged to the web is broadening rapidly, and it is past the capacities of any human to channel through everything to see which media impacts our assurance. In perspective of an examination of web-based systems administration in Flickr, Locr, Facebook, and Google+, we discuss security proposals and the ability of the creating example of a geo-named online person-to-person communication. Writers gave a thought which customers can stay instructed about which parts of the social Big Data deluge are critical to them.

4 Open Research Challenges and Future Perspectives

Currently, Big Data having many research problems in the term of storage, security, processing and privacy, which needs to take care of the same. Considering universal challenges and research issues counted in storage, analysis, capture, process, sharing, visualization, heterogeneity scalability, etc. Generally, social media application domains are link predictions, location oriented mostly interaction analysis, client interaction and analytics, etc. Research zones are characterized as relocating social media environments, growth of online languages, etc. Divergent research areas of Big Data with several domains like IoT, quantum computing, bio-inspired computing, etc.

Recently, machines are getting in on the demonstration to control endless self-ruling devices by means of web and make the Internet of Things (IoT). Subsequently, machines are turning into the client of the web, much the same as people with the web programs. Internet of Things (IoT) is pulling in the consideration of late specialists for its most encouraging open doors also, challenges. Difficult challenge in IoT is the acquisition of information from dissimilar sources. Machine learning techniques and algorithm offer the infrastructures to develop

drksujatha23@gmail.com

frameworks to handle this situation [1]. Bio-inspired algorithms are inspired by nature to minimize the real-world complex problems and provide minimization and optimization techniques for them. Mostly, these techniques are derived by bio-molecules such as protein structure to define complicated calculations to process the data. Breaking down this information also, arranging for content, picture and video, etc. will require parcel of insightful investigation from information researchers and huge information experts, which is a real challenging task [1, 13]. Security challenges are needing to take care of unstructured data on cloud a plausible method shown in [12].

5 Conclusion

This paper summarizes a general plan concerning Big Data in context of social media. General techniques accustomed method Big Data additionally lined up during this paper. The review enlightens Big Data application areas and its contributions to social media analysis. Further, the work and analysis done by many researchers are represented in the paper. Finally, this paper focuses on completely different open analysis difficulties and future perspective of Big Data in several areas.

References

1. Acharjya DP (2016) A survey on big data analytics: challenges, open research issues and tools. Int J Adv Comput Sci Appl (IJACSA) 7(2):511–518
2. Cheung M (2017) An analytic system for user gender identification through user shared images. ACM Trans Multimedia Comput Commun Appl (TOMM) 13(3):30
3. DataFloo (n.d.). https://datafloq.com/read/4-benefits-public-sector-governments-start-big-dat/171
4. Hameed K (2017) Today's social network sites: an analysis of emerging security risks and their counter measures. In: 2017 international conference on communication technologies (ComTech). IEEE pp 143–148
5. Immonen AP (2015) Evaluating the quality of social media data in big data architecture. IEEE Access 3:2028–2043
6. Lee KA (2017) Forecasting influenza levels using real-time social media streams. In: 2017 IEEE international conference on healthcare informatics (ICHI). IEEE, pp 409–414
7. Liu QN (2017) Big data for social media evaluation: a case of WeChat platform rankings in China. In: 2017 IEEE second international conference on data science in cyberspace (DSC). IEEE, pp 528–533
8. Nambisan PL (2015) Social media, big data, and public health informatics: ruminating behavior of depression revealed through twitter. In: 2015 48th Hawaii international conference on system sciences (HICSS). IEEE, pp 2906–2913
9. Patranabish D (2016) Your story. https://yourstory.com/2016/11/big-data-impacting-e-commerce-industry/

drksujatha23@gmail.com

10. Smith MS (2012) Big data privacy issues in public social media. In: 2012 6th IEEE international conference on digital ecosystems technologies (DEST). IEEE, pp 1–6
11. Subasinghe KD (2015) A big data analytic identity management expert system for social media networks. In: 2015 IEEE international WIE conference on electrical and computer engineering (WIECON-ECE). IEEE, pp 126–129
12. Tripathi D (2016) Model for heterogeneous data integration on cloud. In: 2016 3rd international conference on computing for sustainable global development (INDIACom). IEEE, pp 1307–1309
13. Wang L (2014) Bio-inspired cost-effective access to big data, p 243
14. Yadav AY (2016) A secure approach of image encryption using QR code on social media. In: 2016 3rd international conference on computing for sustainable global development (INDIACom), pp 1126–1129

drksujatha23@gmail.com

A Review Paper on Eye Disease Detection and Classification by Machine Learning Techniques



Neha Bharti, Geetika Gautam and Kirti Choudhary

Abstract The regularly expanding measures of patient's information as medicinal image, forces new difficulties to clinical routine, for example, diagnosis, treatment and checking. Image mining is the way toward seeking and finding profitable data learning of information. It is connected on image preparing and machine learning. Picture handling is having essentially for ailment recognition on restorative pictures. These sickness acknowledgment and arrangement are particular to human organ and image nature. With help of image processing and machine learning strategies it is conceivable to computerize as well as help doctors in clinical analysis. In this paper depicts the utilization of different image processing and machine learning methods for identification of eye diseases.

Keywords Medical image mining · Image processing · NB · KNN SVM · AUC · DCT · HMM and PCA approaches

1 Introduction

Lots of individuals in rural and semi urban areas get suffered from eye diseases such as Diabetic Retinopathy, Glaucoma; Age primarily based Macular Degradation and etc. Here victimization Kyrgyzstani monetary unit strategies and technique takes symptoms and take image of illness eye into thought and can discover and classify. Using this we are able to minimize the requirement of the doctor and it will conjointly apprise the patient concerning his illness and its solution. These preparing stages are:

N. Bharti · G. Gautam (✉) · K. Choudhary
Jaipur Engineering College and Research Center, Jaipur, India
e-mail: geetikagautam.cs@jecrc.ac.in; geetikagautam16@gmail.com

N. Bharti
e-mail: nehabharti.it@jecrc.ac.in

K. Choudhary
e-mail: kirtichoudhary.cse@jecrc.ac.in

© Springer Nature Singapore Pte Ltd. 2019

V. S. Rathore et al. (eds.), *Emerging Trends in Expert Applications and Security*,
Advances in Intelligent Systems and Computing 841,
https://doi.org/10.1007/978-981-13-2285-3_74

drksujatha23@gmail.com

1. **Image Processing**: Various picture handling methods utilized as a part of mechanized late determination and investigation of different eye sickness are Enhancement, Registration, Image Fusion, picture Segmentation, Feature extraction [1], design coordinating, grouping, Statistical estimations and analysis [2].

   - **Enhancement**: The image enhancement techniques are most important for medical imaging which can easily affect by noise impedance and different elements that influence the image. Improvement in the quality can be accomplished by utilizing some basic strategies which are recorded beneath: Noise suppression, Sharpening, Contrast Enhancement, Image Segmentation, Feature extraction, Statistical analysis, and Classification based on a classifier. These techniques are helps in improving the quality of the image and algorithms used in these methods are depends upon that condition or situation.
   - **Image Recognition**: The objective of image recognition is the identify the important or basic portion of pictures. Picture characterization includes highlight identification property estimation; picture depiction includes, likewise, division and social structure extraction [3]. Some critical thoughts in each of these territories are checked on in the accompanying sections. Verifiable, the procedures utilized have for the most part been produced on heuristic grounds, however there is expanding enthusiasm for determining ideal methods in view of models for the classes of pictures to be dissected.

2. **Machine learning**: Machine learning is utilized the machines to show, how to deal with the input information with more proficiently. Now and again in the wake of survey the information, we can't translate the example or concentrate data from the information. All things considered, we apply machine learning [4]. With the great quantity of datasets accessible, the interest for machine learning is in rise. Numerous enterprises from drug to military apply machine figuring out how to separate pertinent data. The motivation behind machine taking in is to gain from the information. Numerous investigations have been done on the most proficient method to influence machines to learn independent from anyone else [5, 6]. Many mathematicians and researchers apply a few ways to deal with discover the arrangement of this issue. Some of the techniques of machine learning are explained as;

Naive Bayes (NB): This classifier has been generally and effectively connected for examine on medicinal data [3]. This technique is one of the very compelling and productive characterization calculations, through examination of NB with other well-known classifiers, like Logistic relapse, closest neighbour, Decision Tree, and Rule Based on medicinal informational indexes. The Classifiers are looked at relying upon the region under the Receiver Operating Characteristics (ROC) [3] curve [7]. Kononenko (2001) considered NS as a benchmark calculation that in any restorative space must be attempted before some other propelled strategy. Contrasted with different classifiers, Naive Bayes is straightforward,

drksujatha23@gmail.com

computationally productive, requires moderately little information for preparing, require not to have part of parameters and it is normally hearty to inaccessible and noisy information.

K-Nearest Neighbour (KNN): This is a sort of case based learning, where the capacity is just privately approximated and all calculation is referred until classification [3]. This procedure is called lethargic learning since, it needn't bother with any preparation or negligible preparing stage. All the preparation information is required just amid the testing stage and this system utilizes all the preparation information so that on the off chance that we have an expansive informational collection then we require uncommon strategy to chip away at part of information which is the algorithmic approach [3]. In spite of the fact that characterization is the essential utilization of KNN, we can likewise utilize it for thickness estimation moreover. The k-closest neighbour calculation is one of the most straightforward calculations of all machine learning algorithms. KNN classification was planned from the necessity to play out a few examinations when dependable parametric assessments of likelihood densities are not known or hard to decide [8].

Support Vector Machine (SVM): In machine learning support vector machines (SVMs otherwise called Support Vector Networks) are supervised learning models with correlated learning calculations that learn information and decide designs, utilized for regression and classification analysis [3]. Given an arrangement of preparing cases, each set apart as referring to one classification for one of two classes, a SVM preparing calculation makes a model that partitions new cases into one class or the other contriving it a non-probabilistic parallel direct classifier [3]. A SVM demonstrate is a portrayal of the case as focuses in space allocated with the goal that cases of the diverse classifications are separated [3]. Notwithstanding performing straight characterization, SVMs can quickly play out a nonlinear grouping utilizing the trap called the bit trap, verifiably mapping their into high-dimensional element spaces.

HMM (Hidden Markov Model): We represent an installed HMM [3]-based approach for confront acknowledgment and identification that uses a viable arrangement of perception vectors picked up from the 2D-DCT coefficients. The installed HMM can form the two dimensional information better than the one-dimensional HMM and is computationally less difficult then the two-dimensional HMM [3]. It is well suited for face images since it utilizes important facial characteristics, structure of "states" inside each of that "super states".

DCT: In the part of images processing and recognition, discrete cosine change (DCT) [3] and direct separation are two broadly utilized systems. In view of them, we display another face and palm print acknowledgment approach in this paper. It first uses a two-dimensional distinguishableness judgment to choose the DCT [9] recurrence groups with positive straight distinctness. Chosen groups, it extricates the direct discriminative highlights by an enhanced Fisherface strategy and play out the arrangement by the closest neighbour classifier [3]. We point by point break down hypothetical focal points of our approach in highlight extraction. It can

drksujatha23@gmail.com

fundamentally enhance the acknowledgment rates for confront and palm print information and adequately lessen the measurement of highlight space.

PCA: Another system instituted two-dimensional segment investigation is improved the situation image portrayal. Instead of PCA, 2DPCA [3] depends on image frameworks as opposed to 1D factor so; the image measurements does not should be changed into a factor preceding for highlight extraction. Rather a image covariance measurements is built specifically utilizing the first image networks and its eigenvector are determined for image highlight extraction [3]. To test 2DPCA and assess the execution, a progression of investigations were performed all over image databases: ORL, AR and Yale confront databases. The Experimental outcome demonstrates that the extraction of picture highlights is computationally extremely proficient utilizing 2DPCA than PCA.

AUC: The AUC [3] is the a piece of performing matric of calculated relapse is a generally for utilized assessment matric for twofold characterization issues, such as foreseeing an illness is there or not.

The remainder of this paper is organized as follows: Sect. 2 discusses the literature research; Sect. 3 presents the proposed multistep system, and section presents the concludes our study and future enhancement os this proposed system.

2 Literature Survey

Image preparing is having is importance for infection on restorative pictures. Postulations illness acknowledgment and characterization are particular to human organ and picture write. In The paper has characterized think about on ailment acknowledgment methodologies, for example, SVM, DCT, HMM, and PCA approaches. This paper likewise characterizes the picture preparing operation connected to channel the medicinal picture and to perform infection territory division. To perform picture preparing and sickness location, a progression of picture handling operations are required to enhance the nature of procured picture and to play out the identification [3].

This review paper depicts the utilization of different picture preparing methods for programmed identification of glaucoma. Glaucoma is a neurodegenerative issue of optic nerve, which causes incomplete loss of vision. Substantial number of individuals experiences eye malady depends after looking at retinal fundus picture combination, picture division, highlight extraction, picture improvement, morphology, design coordinating, picture grouping, investigation and factual estimations [2].

In this paper, a novel approach for programmed order of fundus pictures is proposed. The strategy utilizes picture and information pre-preparing strategies to enhance the execution of machine learning classifiers. Promote a discretization technique is proposed to enhance the exactness of the classifiers. Trials were done

drksujatha23@gmail.com

on retinal fundus pictures utilizing the proposed technique on three classifiers Naive Bayes NB, k closest neighbour KNN and bolster vector machine SVM. Results as far as exactness of characterization and region under ROC bend AUC demonstrate that NB beat alternate classifiers according to the proposed strategy [7].

In this exploration a therapeutic picture order system utilizing information mining methods is proposed. It includes the component extraction, highlight determination, and include discretization and characterization. In the order stage, the execution of the customary KNN k-closest neighbour classifier is enhanced utilizing an element weighting plan and a separation weighted voting rather than basic larger part voting. Highlight weights are ascertained utilizing the intriguing quality measures utilized as a part of affiliation administer mining. Trials on the retinal fundus pictures demonstrate that the proposed structure enhances the arrangement precision of customary KNN from 78.57 to 92.85% [8].

The target of our proposed work is to recognize retinal drain for programmed screening of DR utilizing Support Vector Machine (SVM) classifier. To recognize retinal discharge, retinal fundus pictures are taken from Messidor dataset. After pre-preparing, retinal pictures utilizing pixel of same shading and force, the picture is apportioned into non-covering zone that covers the whole picture. Splat and GLCM include are removed to enhance the order exactness. Keeping in mind the end goal to characterize the given information pictures, distinctive classes must be spoken to utilizing applicable and noteworthy highlights with the assistance of determination technique that is prepared by channel and wrapper approaches. At that point discharge influenced retina is distinguished by SVM classifier. At last grouping precision is contrasted and K-Nearest Neighbour (KNN) classifier [10].

Automatic classifiers can be useful for radiologists in distinguishing between benign and malignant patterns. Thus, an artificial neural network (ANN) which might be filled in as a programmed classifier is examined. In medical image process, ANNs are applied to a spread of data-classification and pattern recognition tasks and become a promising classification tool in breast cancer [11].

Image quality can be recognized in numerous perspectives, for example, surface, shading, shape, and spatial relations. They can mirror the unobtrusive variation in numerous degrees. Thus, unique determinations of picture highlights will bring about various grouping choices. These orders can be separated into three kinds: to begin with, the technique in view of measurements, for example, Support Vector Machine; second, the method based on rule, such as decision tree and rough sets; and third, artificial neural network [12].

For cancer detection and classification, image segmentation has been generally utilized. Numerous image segmentation, in light of histogram highlights, edge detection, region developing, or pixel classification, has been prepared utilizing ANNs [13].

drksujatha23@gmail.com

3 Proposed System

The proposed framework will introduce the outline of a specialist framework that expects to give the patient and finding of the eye ailments. The eye has dependably been seen as a passage to the inner workings of the human body structure. Side effects of the eye, in the same way as other of the circumstances there is a swollen eye or red eyes, or redness of eyes. These manifestations can be effortlessly recognized by taking a gander at the eyes.

The proposed system is planning a specialist framework that will take the picture of the patient's eye. The patient will likewise choose or embed alternate indications of the sickness that are experienced by him. A large number of the circumstances it has been watched that, the two distinct maladies can have a similar picture for the sickness however the side effects of the ailment may fluctuate. Along these lines, the framework will take every one of these indications and picture of an eye into thought and will produce the proper outcome that will inform the patient about his sickness. This framework can limit the need of the specialist or it can be utilized where the accessibility of the specialist is less (Fig. 1).

A usage of the picture improvement calculation for the exactness of the picture of the patient's eye. Readiness of the dataset which store different pictures of the illnesses and it will be utilized while testing the genuine information. A usage of the picture examination calculation which will utilize the picture of the eye and the dataset that comprising of the different pictures of ailing eye. Highlight extraction from the picture correlation and testing for the side effects of the sickness entered by the client (Fig. 2).

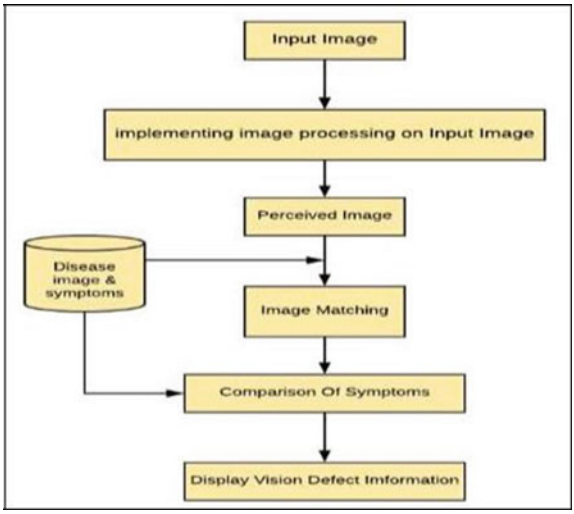


Fig. 1 System architecture

drksujatha23@gmail.com

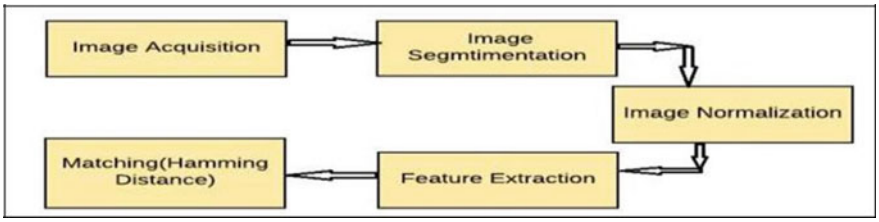


Fig. 2 Image processing techniques [15]

1. **Image Acquisition**

Picture securing is the initial phase in the iris acknowledgment framework. The size and shade of iris of each individual is diverse along these lines it is extremely hard to perceive. The obtaining process [14, 15] produces diverse outcomes for the same people because of the diverse lighting impact, unique situating and distinctive division of separation (Fig. 3).

2. **Image Segmentation** [12]

The Image division is the procedure of devour all the distinctive parts of the eye like student diameter [9], eyelashes, eyelid, sclera, retina part of eye, internal and external piece of the eye and expels every superfluous detail to expand the effectiveness and same time on acknowledgment process. Internal limit and external limit of average iris can be taken as circles. The two circles are normally not to be co-driven (Fig. 4).

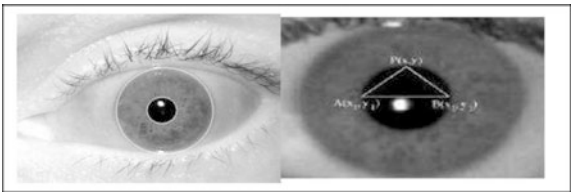
3. **Image Normalization**

Standardization alludes to setting up a division of information picture for the element extraction process. Because of the variety of the enlightenment and the related

Fig. 3 Image acquisition



Fig. 4 Image segmentation



drksujatha23@gmail.com

versatile miss happenings in the iris surface the measure of student may change and may interface with the consequences of example matching [15]. For proposed framework we will utilize Daugman's standardized show [16]. In this model the procedure will create the iris under the different lighting effects.

4. **Feature Extraction**

The information of image highlights are removed by round symmetric channel strategy and grabber channel technique [10]. This technique defines the connection between low recurrence data and high recurrence data [17]. It enhances the effectiveness and accuracy of the eye sickness acknowledgment framework with the assistance of picture pre-handling and highlight portrayal. The inward and external limit cut-off points of an info iris confined by separating, edge recognition and Hough transform [15].

5. **Matching**

This will be the last step for proposed framework. In this progression the encoding procedure will remove the element from iris image and utilized for the coordinating process [15]. The encoding procedure will encode the example of sickness picture into 3002 piece iris code. After the encoding procedure the Hamming Distance strategy will be accustomed to coordinating procedure, this technique gives the measure in no good examples that what numbers of bits are same. The reason for hamming separation decrease the mistakes thought process by false acknowledge and false reject rate.

4 Conclusions and Future Work

The paper has characterized an overview of restorative image preparing and machine learning methods for identifying and ordering eye disease images for illness acknowledgment. Proposed framework will utilize all the image handling methods and calculations said in the paper. Eye ailment discovery and acknowledgment can be accomplished by proposed framework with the utilization of Image Processing and Data Mining procedures.

The proposed system is planning a specialist framework that will take the picture of the patient's eye. A large number of the circumstances it has been watched that, the two distinct maladies can have a similar picture for the sickness so system can enhance to accept disease images along with no of symptoms of the same disease. This will help to provide the better mining result. While capturing the maladies image, it might be possible for addition of noise due to accusation process or dust in device leans. This noisy image will decrease the performance level so de-noise algorithm can be used in the proposed system for the better diagnosis result. And while matching of the images the proposed system can be train and test the images using neural network algorithm for increasing the efficiency of the system.

drksujatha23@gmail.com

References

1. Fu K-S, Rosenfeld A (1976) Pattern recognition and image processing. IEEE Trans Comput C-25(12)
2. Preeti, JP (2013) Review of image processing technique for glaucoma detection. Int J Comput Sci Mobile Comput 2(11):99–105
3. Parul, NS (2015) Study on retinal disease classification and filtration approaches. Int J Comput Sci Mobile Comput 4(5):158–165. ISSN 2320-088X
4. Richert W, Coelho LP (2013) Building machine learning systems with python. Packt Publishing Ltd. ISBN 978-1-78216-140-0
5. Welling M (2011) A First Encounter with Machine Learning
6. Bowles M (2015) Machine learning in python: essential techniques for predictive analytics. Wiley. ISBN 978-1-118-96174-2
7. Mangai JA, Nayak J, Santhosh Kumar V (2013) A novel approach for classifying medical images using data mining techniques. Int J Comput Sci Electron Eng (IJCSEE) 1(2). ISSN 2320–4028 (Online)
8. Mangai JA, Wagle S, Santhosh Kumar V (2013) An improved k nearest neighbour classifier using interestingness measures for medical image mining. World Acad Sci Eng Technol Int J Biomed Biol Eng 7(9)
9. Gupta S, Gagneja A (2014) Proposed iris recognition algorithm through image acquisition technique. Int J Adv Res Comput Sci Softw Eng 4(2). ISSN 2277 128X
10. Inbarathi R, Karthikeyan R (2014) Detection of retinal hemorrhage in fundus images by classifying the splat features using SVM. Int J Innov Res Sci Eng Technol 3(3)
11. Lo S-CB, Chan H-P, Lin J-S, Li H, Freedman MT, Mun SK (1995) Artificial convolution neural network for medical image pattern recognition. Neural Netw 8(7–8):1201–1214
12. Da C, Zhang H, Sang Y (2015) Brain CT image classification with deep neural networks. In: Proceedings of the 18th Asia Pacific symposium on intelligent and evolutionary systems, vol 1, pp 653–662
13. Tang H, Tan KC, Yi Z (2007) Competitive neural networks for image segmentation. Stud Comput Intell 53:129–144
14. Bowyer KW, Hollingsworth K, Flynn PJ (2008) Image understanding for iris biometrics: a survey. Comput Vis Image Underst 110(2):281–307
15. Gupta R, Saini H (2011) Generation of iris template for recognition of iris in efficient manner. Int J Comput Sci Inf Technol 2(4):1753–1755
16. Oad P, Ahmad W (2012) Iris localization using Daugman's algorithm
17. Patil RB, Deshmukh RR (2013) A review on feature extraction techniques of iris. Int J Eng Res Technol (IJERT) 2(12). ISSN 2278-0181

drksujatha23@gmail.com

Kernel FCM-Based ANFIS Approach to Heart Disease Prediction



Waheeda Rajab, Sharifa Rajab and Vinod Sharma

Abstract Heart disease is the main reason for deaths currently in the world. This disease not only affects the old people but also middle-aged and young people. Therefore, the early and precise detection of this disease using intelligent techniques has gained a lot of importance. The goal of this paper is to introduce a diagnostic tool for the detection of heart disease using kernel-based fuzzy C-means clustering, FCM-based adaptive neuro-fuzzy inference system (ANFIS). In the conventional FCM clustering, Euclidean distance is used to compute the distance measure between data points during the clustering process. In kernel-based FCM (KFCM), kernel functions are used to compute this distance measure that enables mapping dataset to high-dimensional space in which data is clearly separable. This generalization helps to make experimental input–output dataset better and distinctly separable leading to more precise data partitions and therefore, more accurate cluster centers. Therefore, these cluster centers when used in fuzzy rule base induction can be used to construct a more precise rule base for the ANFIS which would increase the prediction performance of the system in the analysis of heart disease. For the evaluation of the proposed system, we employed the Cleveland Heart Disease data from the UCI machine learning repository.

1 Introduction

As per the study conducted by India Today, heart disease has emerged as the major cause behind deaths in both rural and urban areas of India. The heart diseases are also called cardiovascular diseases and affect heart and blood vessels. The primary

W. Rajab (✉) · S. Rajab · V. Sharma
University of Jammu, Jammu, India
e-mail: waheedarajab@yahoo.com

S. Rajab
e-mail: sharifa18mca@gmail.com

V. Sharma
e-mail: vnodshrma@gmail.com

© Springer Nature Singapore Pte Ltd. 2019

V. S. Rathore et al. (eds.), *Emerging Trends in Expert Applications and Security*,
Advances in Intelligent Systems and Computing 841,
https://doi.org/10.1007/978-981-13-2285-3_75

drksujatha23@gmail.com

cause of these diseases is atherosclerosis, which is a process related to the building up of a substance called plaque in the walls of the arteries. The early and correct diagnosis of this disease in its early stage is important in decrease the death rate due to this disease. With the advent of computer technology, the use of intelligent methods and algorithms (e.g., artificial-intelligence-based techniques) is playing a crucial role in the diagnosis of complex and uncertain medical problem of disease diagnosis. Recently various studies [1–6] have been conducted on the use of various artificial intelligence (AI) technologies like genetic algorithms, artificial neural networks (ANN), fuzzy systems, and hybrid systems based on two or more AI techniques for accurate detection of this disease. The hybrid AI techniques, in particular, are indispensable in this field as standalone AI techniques are not capable of capturing the inherent uncertainty in the heart disease diagnosis which is based on numerous parameters.

Hybrid neuro-fuzzy systems have also been widely used in this field, the most prominent of which is ANFIS [7] which is based on the concept of fuzzy inference systems and neural networks that use learning to fine-tune its fuzzy rule base for optimizing the system inference process. Data clustering is a useful method for fuzzy modeling where the data clusters obtained from the experimental data using clustering are used to produce fuzzy rules for ANFIS. Fuzzy identification using clustering is a process consisting of finding clusters in the data space and subsequently using the obtained cluster centers to find the premises and consequents of the fuzzy rules. Therefore, the accuracy of the clustering process determines the quality of the rule base and hence the performance of the resulting fuzzy model. The FCM clustering technique was proposed by Dunn [8] and was later improved by Bezdek et al. [9]. It is one of the most popular techniques employed for clustering data sets and an efficient methodology used in fuzzy modeling.

Recently, the kernel methods gained popularity in handling different classification and regression problems. Kernel methods have the ability to improve the precision of computations by shifting data onto a high-dimensional space using kernel functions. The accuracy of clustering using FCM is also improved by using kernel functions in calculating the distance measure between data points during the clustering process. This results in more precise cluster centers [10]. Higher clustering accuracy is achieved because the kernel-induced distance measure increases the data separability by using a high-dimensional space. The KFCM, therefore can be employed to build a more useful rule base for ANFIS which would improve the overall prediction accuracy of the system. In the current paper, we propose a novel diagnostic system for the detection of heart disease by employing ANFIS modeling technique based on KFCM. To construct this prediction model, the experimental data set consisting of various parameters involved in the detection of heart disease is first partitioned into clusters using FCM technique. The resulting cluster centers are then subsequently used to build the initial fuzzy rule base for ANFIS and then the resulting rule base is optimized using a hybrid parameter tuning algorithm composed of least square estimation and gradient descent method. The effectiveness of the KFCM-based ANFIS model has been tested on Cleveland heart disease dataset retrieved from the UCI repository at http://archive.ics.uci.edu/ml.

drksujatha23@gmail.com

A comparison with ANFIS based on conventional FCM has also been presented and discussed.

This paper organization is as follows. In Sects. 2 and 3, the research methodologies have been provided that gives the details of the KFCM algorithm and provides an overview of ANFIS. Section 4 presents the experimental dataset. Section 5 discusses the experimental results, a comparison of performance with the ANFIS based on conventional FCM and ANN has been given and Sect. 6 provides the concluding remarks on this study where various enhancements to this work have also been presented.

2 Kernel FCM Clustering

The KFCM was proposed by Qiang et al. [10] as an enhancement of the standard FCM clustering algorithm based on the use of kernel functions. For a dataset $X = \{x_1, x_2, \ldots, x_n\}$, the conventional FCM algorithm calculates the fuzzy subsets of X by optimizing the following objective function:

$$\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^{m}\left||x_j - v_i\right||^2 \tag{1}$$

where c is the number of cluster centers, n is the number of data instances, μ_{ij} is the membership of x_j in ith class, v_i is ith cluster center, $\left||x_j - v_i\right||^2$ is the measure of distance and parameter m is used to limit the fuzziness of clustering. In KFCM, the distance measure is generalized by employing a nonlinear mapping $\emptyset$ from input space to a higher dimensional space, i.e.,:

$$\emptyset : x \rightarrow \emptyset(x)$$

Using this nonlinear mapping, the dot product '$x_i.\ x_j$' is used as a similarity measure in various learning algorithms that can be mapped to a more general measure: $\emptyset(x_i) \cdot \emptyset(x_j)$. This dot product in higher dimensional space is calculated using a kernel function $K(x_i, x_j)$, i.e.:

$$\emptyset(x_i) \cdot \emptyset(x_j) = K(x_i, x_j) \tag{2}$$

The distance measure $\left||x_j - v_i\right||^2$ in input space in terms of function $\emptyset$ therefore is given by

$$\left||x_j - v_i\right||^2 = \left||\emptyset(x_i) - \emptyset(x_j)\right||^2 \tag{3}$$

drksujatha23@gmail.com

where

$$\begin{aligned}\left|\left|\emptyset(x_j)-\emptyset(v_i)\right|\right|^2 &= \left(\emptyset(x_j)-\emptyset(v_i)\right)-\left(\emptyset(x_j)-\emptyset(v_i)\right)\\ &= \emptyset(x_j).\emptyset(x_j)-2\emptyset(x_j)\emptyset(v_i)+\emptyset(v_i).\emptyset(v_i)\\ &= K(x_j,x_j)-2K(x_j,v_i)+K(v_i,v_i)\end{aligned} \tag{4}$$

Therefore, using (3) by the kernel approach, the objective function in KFCM is given by

$$J_m(U,V)=\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^{m}\left|\left|\emptyset(x_j)-\emptyset(v_i)\right|\right|^2 \tag{5}$$

From Eq. (4),

$$\left|\left|\emptyset(x_j)-\emptyset(v_i)\right|\right|^2=K(x_j,\ x_j)-2K(x_j,v_i)+K(v_i,v_i) \tag{6}$$

K(x, y) can be any kernel function, for example, Gaussian kernel, polynomial kernel, Fisher kernel, etc. Using Eq. (12), (11) becomes

$$\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^{m}\left(K(x_j,x_j)-2K(x_j,v_i)+K(v_i,v_i)\right) \tag{7}$$

Gaussian function is a common kernel function given by

$$K(x,y)=e^{\left(-||x-y||^2/\sigma^2\right)} \tag{8}$$

where $K(x,\ x)=1$ and σ is an adjustable parameter. Using the Gaussian kernel function, Eq. (7) becomes

$$\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^{m}\left(1-K(x_j,v_i)\right) \tag{9}$$

where

$$\mu_{ij}=\frac{\left(1/\left(1-K(x_j,v_i)\right)\right)^{1/(m-1)}}{\sum_{k=1}^{c}\left(1/\left(1-K(x_j,v_k)\right)\right)^{1/(m-1)}} \tag{10}$$

$$v_i=\frac{\sum_{j=1}^{n}\mu_{ij}^{m}K(x_j,v_i)x_j}{\sum_{j=1}^{n}\mu_{ij}^{m}K(x_j,v_i)} \tag{11}$$

Other kernel functions can also be used so that above equations can be modified accordingly.

drksujatha23@gmail.com

Algorithm for KFCM:

Step 1: set k=0, m > 1 and $\epsilon > 0$.
Step 2: set the initial memberships values μ_{ij}^{0}.
Step 3: I) Update all v_i^k using eq. (11).
II) Update all μ_{ij}^k using eq. (10).
If $\max\left(\left|\mu_{ij}^k - \mu_{ij}^{k-1}\right|\right) <= \epsilon$ Stop
else k=k + 1
go to step 3.
end if

3 ANFIS

ANFIS is an adaptive system that has the learning capability to optimize the performance based on finding the best parameters for the fuzzy rules within its rule base. Figure 1 shows ANFIS structure having inputs x_1 and x_2 and a rule base consisting of consisting of two Sugeno-type fuzzy rules:

If x_1 is A_1 and x_2 is B_1 then $f_1 = p_1x_1 + q_1x_2 + r_1$
If x_1 is A_2 and x_2 is B_2 then $f_2 = p_2x_1 + q_2x_2 + r_2$

A detailed architecture of ANFIS is provided as follows.

Layer 1: This layer receives the input and consists of nodes with adaptive node functions. Each node has an output equal to

$$O_{1,i} = \mu A_i(x) \quad \text{for } i = 1, 2 \tag{12}$$

The output of each node is the value of the membership function A of that node and $O_{k,I}$ is the node in the ith position of the kth layer.

Various types of membership function can be used, like Gauss function, the bell-shaped function, etc.

Layer 2: Each node in this layer computes the product of incoming signals with output given by

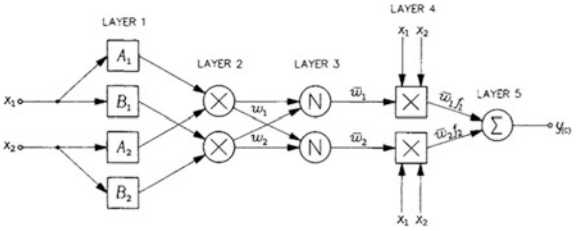


Fig. 1 ANFIS architecture

drksujatha23@gmail.com

$$O_{2,i} = w_i = \mu A_i \mu B_i(y),\ i = 1, 2 \tag{13}$$

Layer 3: In this layer, each jth node computes the ratio of the firing strength of the jth rule to the summed firing strengths related to all other fuzzy rules, with output:

$$O_{3,j} = w_j = \frac{w_j}{w1 + w2},\quad j = 1, 2 \tag{14}$$

Layer 4: In this layer function for ith node is

$$O_{4,1} = \bar{w}_i f_i = \bar{w}_i(p_i x_1 + q_i x_2 + r_i) \tag{15}$$

Layer 5: In this layer, a single node is used to calculate the overall output of ANFIS as the total sum of all the output signals from the previous layer:

$$O_{5,1} = \sum_i \bar{w}_i f_i = \frac{\sum_i w_i f_i}{\sum_i w_i} \tag{16}$$

where $O_{5,1}$ is the obtained output available to the user.

For the optimization of the fuzzy rule base of ANFIS, either standard back-propagation or the hybrid learning algorithm can be used. The hybrid learning is mostly used and is an effective technique that employs gradient descent technique to tune the premise parameters of the fuzzy rules and LSE is used to find the optimal consequent parameters.

4 Dataset

The heart disease dataset used in this study is the Cleveland dataset that was obtained from the online UCI machine learning repository. A number of research studies have used this dataset particularly for classification based on fuzzy logic. The information contained in the dataset consists of 13 attributes (listed in Table 1), which are effective factors in the prediction of heart disease.

5 Experimental Results

The data set for experimentation was divided randomly into training, checking, and testing data, which comprised of 75%, 15%, and 15% of the total Cleveland data, respectively. The training dataset was used as input to a KFCM algorithm for producing the initial rule base for ANFIS which resulted in 10 fuzzy rules and 10 membership functions per input variable. For each of the input variables, Gaussian

drksujatha23@gmail.com

Table 1 Prediction accuracy on test data

Attribute	Abbreviation
Age of the patient in years	age
Sex (1 for male and 0 for female)	sex
Type of chest pain	cp
Blood pressure during resting (in mm Hg)	trestbps
Cholesterol serum in mg/dl	chol
Fasting blood sugar > 120 mg/dl	fbs
Maximum heart rate achieved	thalach
Resting electrocardiographic results	restecg
ST depression due to exercise	oldpeak
Angina due to exercising	exang
Slope of peak exercise ST segment	slope
3 for normal, 6 for fixed defect, and 7 for reversible defect	thal
Number of major vessels	ca

Table 2 Prediction accuracy on test data

Model	Accuracy in %
KFCM-based ANFIS	86
FCM-based ANFIS	84.52
Multilayer ANN	79.01

membership functions were used. The parameters of the system were then fine-tuned using a hybrid learning technique to improve the accuracy of prediction. The training was performed for 1000 epochs but the checking error stabilized after 462nd epoch and the system was saved for a testing purpose. In Table 2, the precision of the KFCM-based ANFIS using Gaussian kernel function on independent test data. Table 2 also shows the forecasting performance of ANFIS based on conventional ANFIS and ANN. It is clear from the results that due to the application of KFCM the accuracy of ANFIS for the detection of heart disease on this dataset can be improved.

References

1. Wikipedia (n.d.) EasyChair. https://en.wikipedia.org/wiki/EasyChair
2. Bezdek JC, Ehrlich R, Full W (1984) FCM: the fuzzy C-means clustering algorithm. Comput Geosci 191–203
3. Zhang Song D-Q, Chen SC (2004) A novel kernelized fuzzy C-means algorithm with application in medical image segmentation. Artif Intell Med. https://doi.org/10.1016/j.artmed.2004.01.012
4. Dunn JC (1974) Some recent investigations of a new fuzzy partition algorithm and its application to pattern classification problems. J Cybern pp. 1–15

drksujatha23@gmail.com

5. Jang J-SR (1993) ANFIS: adaptive-network-based fuzzy inference system. IEEE Trans Syst Man Cybern 665–685
6. Hsieh N-C, Hung L-P, Shih C-C, Keh H-C, Chan C-H (2012) Intelligent postoperative morbidity prediction of heart disease using artificial intelligence techniques. J Med Syst 1809–1820
7. Voronkov A, Hoder K (n.d.) Templates. https://easychair.org/proceedings/template.cgi?a=12732737
8. Voronkov A (2014) Keynote talk: EasyChair. In: Proceedings of the 29th ACM/IEEE international conference on automated software engineering, pp 3–4. ACM
9. Carlisle D (2010) Graphicx: enhanced support for graphics. http://www.ctan.org/tex-archive/help/Catalogue/entries/graphicx.html
10. Voronkov A (2004) EasyChair conference system. easychair.org

drksujatha23@gmail.com

State-of-the-Art Artificial Intelligence Techniques in Heart Disease Diagnosis



Nahida Nazir, Sharifa Rajab and Vinod Sharma

Abstract Artificial intelligence (AI)-based techniques are gaining tremendous importance and popularity in designing medical diagnostic systems. In this context, various AI techniques like decision trees, Naïve Bayes networks including various soft computing techniques like artificial neural networks (ANN), genetic algorithms, and fuzzy-logic-based systems have been predominantly applied in the diagnosis of heart diseases. This paper examines the recent advances in the research based on the application of some of the popular AI techniques in heart disease prediction. Along with each technique the main strengths, limitations, and future directions have been presented.

1 Introduction

Heart and blood vessel diseases also called Cardiovascular diseases lead to several health problems, which are mainly result of the process called atherosclerosis. Atherosclerosis is a condition which is caused by the effect of building up of a substance called plaque in the walls of arteries [1]. Due to this reason, the arteries of the person start shrinking which causes problems for the flow of blood through them. As a result, this increases the potential risk of causing heart disease in a person. According to the World Health Organization, heart diseases cause more deaths per year as compared to other ailments. It has been estimated that cardiovascular disease results in more than 80% deaths in middle- and low-income countries, which are increasing per year [2]. With the increased interest in AI-related research, the application of AI techniques like decision trees, artificial

N. Nazir (✉) · S. Rajab · V. Sharma
University of Jammu, Jammu, India
e-mail: nahidanazir449@gmail.com

S. Rajab
e-mail: sharifa18mca@gmail.com

V. Sharma
e-mail: vnodshrma@gmail.com

© Springer Nature Singapore Pte Ltd. 2019

V. S. Rathore et al. (eds.), *Emerging Trends in Expert Applications and Security*,
Advances in Intelligent Systems and Computing 841,
https://doi.org/10.1007/978-981-13-2285-3_76

drksujatha23@gmail.com

neural networks, genetic algorithms, etc., began to play an important role in solving uncertain and complex medical task of diagnosis of diseases. The medical practitioners are also applying computerized technologies to facilitate disease diagnosis due to a lot of uncertainty in this field [3]. Recently, the literature on the application of AI-based methods in medicine has seen numerous related studies [4–8]. In the field of heart disease diagnosis also, there has been a lot of research on the application of AI techniques. Some of these techniques have been more popular due to being suitable for the complex nature of diagnosis in this field. Additionally, many of the AI techniques have been combined with other techniques to enhance the accuracy of diagnosis.

This paper explores the recent advances in the research on the development of heart disease diagnostic tools using AI techniques. The study is based on relevant research articles in journals from various reputed publishing houses like Elsevier, Springer and various international conferences. The goal of the paper is to give an up-to-date view of the major AI techniques used in the heart disease prediction focusing on their potential benefits and drawbacks.

The paper is organized as follows. Sect. 2 captures the recent advances in the research on heart disease prediction. Then, the applications of ANNs, genetic algorithms, and neuro-fuzzy systems in heart disease prediction along with the main limitations and strengths with respect to this field are given in separate subsections. In the last section, the concluding remarks are given along with the possible future scope in this field.

2 Popular AI Methods in Heart Disease Detection

Numerous research studies have been conducted in recent years focusing on various aspects of heart disease diagnosis on different datasets. Different researchers have used different AI techniques in related research. The various techniques used include Naive Bayes networks, ANNs, evolutionary algorithms, support vector machines, fuzzy systems but some of these techniques have been more popular in this field. Also, the hybrid AI systems based on the combination of two or more AI techniques like neuro-genetic systems, neuro-fuzzy systems have become popular in this field. In the following subsections, the recent advances in the research are based on the three important AI techniques predominantly employed in developing diagnostic systems for prediction of heart diseases are presented.

2.1 Artificial Neural Networks in Heart Disease Prediction

ANN is an information processing system inspired by biological nervous systems like a brain in living beings. The main component of this prototype is an innovative system for processing information composed of various highly interconnected

drksujatha23@gmail.com

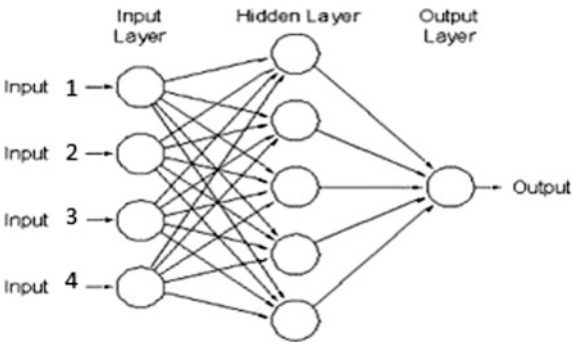


Fig. 1 Structure of a feed-forward ANN

elements called neurons that work in harmony to solve a particular problem. ANNs have the capability of deriving meaning from complex and imprecise data which is difficult to extract using other computational techniques. These models do not need prior knowledge about the domain. Figure 1 shows a typical multilayer ANN system. Due to these benefits, the potential of ANNs in developing diagnostic systems for heart disease detection has been investigated recently in a number of research studies.

In a significant study, Das et al. [9] designed a medical diagnostic system based on ANN ensemble for diagnosis of heart disease. The authors used SAS base software 9.1.3. This ensemble-based model created new models by integrating the values of the predicted output of a number of predecessor models. This resulted in a more effective diagnostic model with 89.01% classification accuracy. In another study, Hsieh et al. [10] devised an ensemble-based medical diagnosis tool for predicting postoperative morbidity after endovascular aneurysm repair. The ensemble was based on support vector machine, Bayesian network, and ANN as basic models. Sonawane and Patil [11] developed a multilayer ANN-based model for prediction of heart disease. The model was based on 13 features as input variables and was trained by backpropagation learning algorithm with a single output which indicated the absence or presence of heart disease in a patient with high accuracy of 98% which was comparative to other systems. Olaniyi and Oyedotun [12] designed an intelligent system based on support vector machine and feed-forward multiple layer ANN. The datasets consisted of various features of patients to be diagnosed with heart diseases and were obtained from UCI machine learning repository. The diagnostic system was used to predict if heart disease was present or not in the patient. ANN showed the accuracy of 87.5% and support vector machine showed the accuracy of 85%. Recently, in a study, Samuel et al. [13] used ANN to predict the heart failure risk. The authors used 13 heart-failure-related attributes and their individual contributions were obtained by consulting an experienced cardiac clinician. Fuzzy analytic hierarchy process was used to compute the global weights for the attributes based on their individual contribution. The experimental result showed that the proposed method achieved an

drksujatha23@gmail.com

average prediction accuracy of 91.10%, which was 4.40% higher as compared to that of the conventional ANN.

As is evident, ANNs have been successfully applied in designing heart disease diagnostic systems but ANNs face some important limitations which need to be addressed. These limitations are mainly related to the more training time needed to build ANN-based diagnostic system due to a number of parameters involved, need of a large dataset for building an ANN, difficulty in interpreting the prediction results due to the black box nature of ANNs. To overcome the training time-related issue, a lot more research is needed in the field of incremental ANN learning. The black box nature of ANNs poses problems as usually human understandable medical diagnostic models are required.

2.2 *Genetic and Evolutionary Algorithms in Heart Disease Prediction*

Genetic algorithms [14] and evolutionary algorithms [15] are basically searching algorithms inspired from the phenomena of natural selection. These are initiated using some of the potential solutions to the problem called the population. From this initial population, solutions are taken and are then used to produce a new population of solutions, which are nearer to the optimal solution to the problem at hand. The basic idea used in this technique is the survival of the fittest. The main strength of these algorithms is that these can better follow the goals and preferences of the user, as these techniques do a global search of the solution space and are able to handle better the attribute interaction than other AI-based algorithms, in which search is more or less local. Genetic algorithms and evolutionary systems have also been popular in the field of heart disease prediction.

In 2012, Jabbar et al. [16] introduced an efficient classification technique using the genetic algorithm technique for the prediction of heart disease. The study was motivated by the fact that by employing a genetic algorithm for discovering forecasting rules is that the rules discovered are very comprehensible, have high interestingness values and high predictive precision. The authors showed that the classifier rules extracted from the model help in the highly accurate prediction of heart disease helpful to doctors in disease diagnosis. In another study, Niranjana Devi and Anto [17] proposed an evolutionary and fuzzy-logic-based expert system for the purpose of diagnosis of the Coronary Artery Disease. The study used the Cleveland clinic foundation datasets for heart diseases. The authors used decision trees for the detection of most significant attributes and the output was then converted into fuzzy rules which were then tuned using a genetic algorithm. The proposed system's forecasting performance was compared to the existing methods for this application. In the same year, Kumar and Sahoo [3] combined Naive Bayes and genetic algorithm for effective classification of patients into those with heart diseases and those without these. The results from the experiments showed that the

drksujatha23@gmail.com

proposed system was effective in the correct classification of patients for this disease. Recently, Long et al. [18] designed a medical diagnostic system based on evolutionary genetic hybrid algorithms, chaos firefly, and interval type-2 fuzzy system. The authors investigated the application of rough-sets-based attribute reduction using chaos firefly algorithm for finding optimal attribute reduction which is helpful in reducing the computational cost and improves the performance of the prediction system. Experimental results demonstrated that the proposed system significantly dominates Naive Bayes, support vector machines, and ANN in terms of forecasting accuracy. More recently, Paul et al. [19] proposed a genetic-algorithm-based fuzzy decision support system for forecasting the risk level of heart disease. The authors addressed various issues related to this prediction problem, i.e., dataset preprocessing, effective attribute selection and weighing of the fuzzy rules. The effectiveness of the proposed system was evaluated on real data sets.

Although genetic algorithms and evolutionary algorithms are suitable for exploring complex, vast and multidimensional problem spaces these have some limitations due to the high computational complexity and less potential for dynamic models.

2.3 Neuro-Fuzzy Systems in Heart Disease Prediction

Neuro-Fuzzy Systems use the learning techniques inspired by ANNs for learning and fine-tuning the parameters of the membership functions obtained from experimental data to be used in a Fuzzy Inference System. Using this hybrid approach, the drawbacks of ANNs and fuzzy systems, i.e., the black box character of ANNs and the problem of obtaining suitable parameters of membership functions in case of fuzzy systems are eliminated. Neuro-fuzzy systems help to automate the process of integrating the expert or domain knowledge into the fuzzy system. Due to the combined strengths of ANNs and fuzzy systems, neuro-fuzzy systems have been widely used in the diagnosis of heart diseases.

Opeyemi and Justice [1] introduced a study with the goal of designing a neuro-fuzzy system for prediction of a heart attack. The authors designed the system using eight input variables and a single output. The various input variables were heart rate, chest pain type, exercise, age, blood pressure, cholesterol, sex, and blood sugar. The output detected the risk level of patients and was set as very low, low, high, and very high. Later on, Sen et al. [20] developed a two-layer approach based on a neuro-fuzzy approach to predict cardiovascular disease in patients. The first level dealt with the determination of the critical factors for the detection of disease and the second level dealt with the actual prediction of the disease. The data set for the model was obtained from the UCI machine learning repository and was effective in the prediction of heart disease. In another study, Abushariah et al. [21] designed a heart disease diagnosis system in MATLAB using adaptive neuro-fuzzy inference system (ANFIS). The dataset for the purpose was the Cleveland dataset

drksujatha23@gmail.com

for heart diseases present in the MATLAB environment. In addition, the authors investigated the impact of different values of various important parameters on the neuro-fuzzy-based system for selecting the optimal parameters for obtaining the highest performance. The results from the experiments showed that the neuro-fuzzy system outperformed the ANN in terms of prediction accuracy. Kolus et al. [22] used ANFIS for classifying heart rate used in the detection of heart disease. The classification of heart rate, which was also the output of the model was done as very light, light, moderate, and heavy. The authors also considered various variability factors (i.e., physical and physiological differences) in patients. Sagir and Sathasivam [23] proposed a diagnostic tool based on ANFIS and Levenberg–Marquardt algorithm for detection of heart disease in patients. The authors showed that the model predicted the degree of patients' heart disease with reliable and more accurate results. A comparison of results with existing methods on the basis of diagnostic accuracy with the Statlog-Cleveland Heart Disease dataset was done.

3 Conclusion

The paper presented the developments related to the research based on the application of ANN, genetic algorithms and neuro-fuzzy systems in the prediction of heart disease. Also, the benefits and shortcomings of these techniques with respect to the application in this field were presented. Form the literature study related to this field, it is evident that different authors have followed different approaches like effective feature selection methods, taking additional relevant factors in patients into consideration, integrating multiple AI methods, etc., to tackle the difficulties in diagnosis of this disease. Even though standalone AI techniques have been demonstrated to perform satisfactorily in the diagnosis of heart disease but we are of the opinion that the future of the intelligent heart disease diagnostic systems lies in hybrid AI techniques. The reason is that the standalone AI techniques are not capable of capturing the inherent uncertainty in the heart disease diagnosis based on numerous parameters. The use of hybrid AI models combines the strengths of multiple intelligent systems provides a useful framework to address the natural complexity in the diagnosis of heart diseases.

References

1. Opeyemi O, Justice EO (2012) Development of neuro-fuzzy system for early prediction of heart attack. Int J Inf Technol Comput Sci 22–28
2. World Health Organization. http://www.who.org. Accessed 9 Feb 2014
3. Kumar S, Sahoo G (2014) Classification of heart disease using Naïve Bayes and genetic algorithm. Comput Intell Data Mining 269–282
4. Adeli A, Neshat MA (2010) Fuzzy expert system for heart disease diagnosis. In: Proceeding of the international multi conference of engineers and computers scientists

drksujatha23@gmail.com

5. Allahverdi N, Torun S, Sarıtaş İ (2007) Design of a fuzzy expert system for determination of coronary heart disease risk. In: International conference on computer systems and technologies, CompSysTech'07
6. Nazmy TM, El-Messiry H, Al-Bokhity B (2010) Classification of cardiac arrhythmia based on hybrid system. Int J Comput Appl 0975–8887
7. Sikchi SS, Sikchi S, Ali MS (2012) Design of fuzzy expert system for diagnosis of cardiac diseases. Int J Med Sci Public Health 56–61
8. Kumar AVS (2013) Diagnosis of heart disease using advanced fuzzy resolution mechanism. Int J Sci Appl Inf Technol 2:22–30
9. Das R, Turkoglu I, Sengur A (2009) Effective diagnosis of heart disease through neural networks ensembles. Expert Syst Appl 7675–7680
10. Hsieh N-C, Hung L-P, Shih C-C, Keh H-C, Chan C-H (2012) Intelligent postoperative morbidity prediction of heart disease using artificial intelligence techniques. J Med Syst 1809–1820
11. Sonawane JS, Patil DR (2014) Prediction of heart disease using multilayer perceptron neural network. In: IEEE international conference on information communication and embedded systems (ICICES). https://doi.org/10.1109/icices.2014.7033860
12. Olaniyi EO, Oyedotun OK (2015) Heart diseases diagnosis using neural networks arbitration. IJ Intell Syst Appl 75–82
13. Samuel OW, Asogbona GM, Sangaiah AK, Fang P, Li G (2017) An integrated decision support system based on ANN and Fuzzy_AHP for heart failure risk prediction. Expert Syst Appl 163–172
14. Goldberg D (1989) Genetic algorithms in search, optimization, and machine learning. Addison-Wesley, Reading, MA
15. Schwefel HP (1995) Evolution and optimum seeking. Wiley, New York
16. Jabbar MA, Chandra P, Deekshatulu BL (2012) Heart disease prediction system using associative classification and genetic algorithm. In: International conference on emerging trends in electrical, electronics and communication technologies-ICECIT
17. Devi N, Anto S (2014) An evolutionary-fuzzy expert system for the diagnosis of coronary artery disease. Int J Adv Res Comput Eng Technol (IJARCET) 212–224
18. Long NC, Meesada P, Unger H (2015) A highly accurate firefly based algorithm for heart disease prediction. Expert Syst Appl 8221–8231
19. Paul AK, Shill PC, Rabin RI, Akhand MAH (2016) Genetic algorithm based fuzzy decision support system for the diagnosis of heart disease. In: 5th International conference on informatics, electronics and vision (ICIEV). https://doi.org/10.1109/iciev.2016.7759984
20. Sen AK, Patel SB, Shukla DP (2013) A data mining technique for prediction of coronary heart disease using neuro-fuzzy integrated approach two level. Int J Eng Comput Sci 2663–2671
21. Mohammad Abushariah AM, Assal AM, Omar YA, Yousef MM (2014) Automatic heart disease diagnosis system based on artificial neural network (ANN) and adaptive neuro-fuzzy inference systems (ANFIS) approaches. J Softw Eng Appl 1055–1064
22. Kolus A, Imbeau D, Dubé PA, Dubeau D (2016) Classifying work rate from heart rate measurements using an adaptive neuro-fuzzy inference system. Appl Ergon 158–168
23. Sagir AM, Sathasivam S (2017) A novel adaptive neuro fuzzy inference system based classification model for heart disease prediction. Pertanika J Sci Technol 43–56

drksujatha23@gmail.com

Security and Privacy Issues in Big Data: A Review



Priyanshu Jadon and Durgesh Kumar Mishra

Abstract In the current digital world, big data concept is increasing very rapidly. Data is generated in very large volume and in a variety of forms. This large and complex dataset is used by the business organizations for finding their customer needs or insights. Therefore, the security and privacy over the large datasets become too much necessary for the organizations and users. This paper mainly deals with the issues related to big data while storing and processing it. In the proposed architecture clients data is distributed among different Hadoop machines and computation is done through a single machine using a random method and joint computation is performed here that announces the final result to the clients. Therefore, our architecture provides the anonymity of users to maintain the high level of privacy, that means machine who performs computation only knows the data as a whole of all clients and does not know to whom the data belongs and thus the privacy of different user during data processing remains anonymous.

Keywords Big data · NameNode · Honeypot · Attribute-based encryption (ABE) Privacy-preserving data publishing (PPDP) · Privacy · Security Analytics

1 Introduction

Within past few years data increasing very rapidly and it becomes a problem for every organization (whether it is small or big) that how to store and process that data which helps them to find out some valuable insights that are necessary for their business.

P. Jadon (✉)
Shri Vaishnav Vidyapeeth Vishwavidyalaya, Indore, India
e-mail: priyanshujadon1@gmail.com; jadonpriyanshu08@gmail.com

D. K. Mishra
Sri Aurobindo Institute of Technology, Indore, India

© Springer Nature Singapore Pte Ltd. 2019

V. S. Rathore et al. (eds.), *Emerging Trends in Expert Applications and Security*,
Advances in Intelligent Systems and Computing 841,
https://doi.org/10.1007/978-981-13-2285-3_77

drksujatha23@gmail.com

Everything near our surrounding may have an increased amount of big data, some of the applications of big data include healthcare department, government department, online business, traffic management, etc. The enormous and complex amount of data is generated by the users to this application, that are very important for the organization through which they completely analyze the complex datasets that give the information about what user wants and provide the result according to the user's need. Various algorithms and analysis and visualizing techniques are required to find out the beneficial information of the user and this increases the efficiency and overall performance.

The data is generated through the Internet, social media, mobiles, sensors, etc., may be responsible for increasing data. These large amounts of data require a smart technology that helps to store it, process it, and analysis it very easily. The growing volume of unstructured and semi-structured data plays a major role in increasing data. Each and every one produces the amount of data constantly that forms a large data for an organization [1]. Big data is a complex data set, whose size is beyond traditional system to store, compute, and analyze it. The data is not increasing in terms of volume but also in velocity and variety of data is increasing.

Volumes, Variety, Velocity are three Vs of big data [2]. These 3Vs form a complete definition of big data mainly given by IBM. These large and complex data sets are then stored and compute in advance system known as Hadoop which is basically a file system, that allows to store and process any type of data whether, it is structured, semi-structured, and unstructured. Hadoop does not provide the proper security and privacy of data and hence it becomes a challenge for the Hadoop to secure data and maintain the privacy of the user.

The rest of the paper is organized as follows. Section 2 introduces the big data issues. Section 3 gives the overview of architecture that preserves the privacy of the user's data while transferring it to the analytics machine and finally in Sect. 4 conclusion is added.

2 Security and Privacy in Big Data

In the system, sensitive data is coming from different clusters of commodity hardware, hence security and privacy become important when data is traveling from clusters to the environment. Data is increasing day by day very rapidly, in every second huge amount of data is generated and both security and privacy become an issue. Every organization deals with big data and this data are collected from various environments. So, there may be a lot of chances of data breaching and many other types of attacks to occur, therefore we have to fight this challenge and have to get some specific solution.

Security and privacy become a very important concern for all the organization because they can find their user insights that depend on the data they collect, therefore, it is a deal for the organizations to secure their user's data for getting a reliable insight that is beneficial for companies for their future growth. The data is

drksujatha23@gmail.com

available in semi-structured and unstructured format hence it becomes a problem to secure this data. Security mainly concentrates on how to protect the data from malicious attacks and it protects an enterprise or agency. Some of the security problems and their solutions are:

The data is coming from different sources and we are storing it in a Hadoop framework, where no inbuilt security is provided, hence the data is at high risk and we need to secure it. The two major weakness of Hadoop security is accessing data on the cloud and HDFS security. Different mechanisms are implemented in both weaknesses that provide a security to data. Trust mechanism, random encryption algorithm are used that provides a secure access for storing data on cloud [3]. In HDFS, multiple copies of the single file are to be present in different data nodes of the environment and hence if any kind of modification is done by an attacker then it becomes difficult to recover the data. So to prevent the data various kinds of approaches are present like Kerberos, Bull eye Algorithm, and NameNode Approach.

We have to process a large amount of data, therefore, computation is taken in a parallel manner. In the Hadoop system, MapReduce phase is responsible for the processing of big data where the large datasets are divided into a number of chunks. The whole data is to process through mappers and reducers that work in a key-value form. Mapper mainly reads the input data chunks and perform operations on it and through reducer, we get the output. Hence, we have to secure our mapper from vulnerable network attacks as it contains the confidential data [4].

An organization may collect the data from different sources, such as from end devices, so it is necessary that the input must validate that means the data must be trustworthy [4]. We control this issue by monitoring the network that is helpful in catching the intrusion activity. Mainly Security Monitoring architecture is present that helps in analyzing the Honeypot data, HTTP traffic, DNS traffic, IP flow [3].

The other big data security issue is key generation and key management. The data is coming from different servers to the client with the help of certain protocols and the transmission can take place on the behalf of the key generated and shared between them. So, it is necessary to provide security to the data when it is in transmitting state. Quantum Cryptography is a technique that solves this kind of problem [3].

Big Data environment consists of many programming tools, like NoSQL, that was not designed by taking security in mind. For e.g., it does not have any mechanism for user access and for encrypting the data and this may lead to a kind of network vulnerabilities. NoSQL also has some kind of disadvantages as compared to traditional systems as they do not follow ACID properties.

Above we discussed security problems and solutions. Now we are going to discuss some securing points in big data [5].

While transferring the data or accessing the data, we have to follow the Kerberos mechanism that may provide a proper authentication, and protects the data while transferring it on the network by encrypting it. We must implement a tool that is used to maintain the overall workflow of the data in a framework. Oozie is a tool

drksujatha23@gmail.com

available that monitors the overall workflow of data that means, Oozie monitors how the data is going from HDFS to MapReduce and so on.

Various security tools are available like apache sentry, project rhino, etc. Apache Sentry is an open-source tool given by Cloudera, that helps us to achieve authentication process and protecting the huge amount of data that is to be stored in a distributed file system. Project Rhino provides an integrated end to end solution to the ecosystem.

Privacy is the ability to decide what information of an individual goes where. Big data effectively help us to understand the world but the amount of data is increasing very rapidly that increased the privacy breach. Some issues related to the privacy in big data and mechanisms are provided below:

Personal information of the user is easy to identify during the transmission over the internet, therefore we need to provide a kind of privacy so that no unauthorized user sees the information. Various kind of encryption techniques is used here like ABE (Attribute-Based Encryption) that provides a secure communication framework between end-to-end entities. ABE may have certain kinds of drawback like the computational overhead s increases during handling the data of different users and it is very hard to find out the attributes [4, 6]. Some other encryption techniques are Identity-based encryption. Homomorphic-based encryption, etc., are used to protect the privacy of the user. This kind of protection is to be carried out at the time of data generation [7].

The sensitive data is to be stored and processed in allocation is not secure properly that may sometimes lead a data leakage. This may be happening in the storing and processing phase of data. During the storage phase, conventional security mechanism is to be used to protect the privacy of the data that is categorized into four different categories as file-level data security schemes, database-level data security schemes, media-level data security schemes, and application-level encryption schemes. Storage Virtualization is also a kind of mechanism which is used to control the three Vs of big data. During the processing phase, other kinds of techniques or mechanism are to be used that provides a protection to the data when it performs the kind of computation on it like Privacy-Preserving Data Publishing, knowledge extraction, and data anonymization that have further different kinds of approaches [6].

The technologies such as NoSQL does not provide robustness to the data so the proper authorization and authentication need to take place due to which no illegal user can use anyone's private data. Before sending the private information to any organization the user may check that, that particular organization is trustworthy or not. Big data is generated from various devices and there will be chances that anyone can access our data so we have to take control over it by providing a certain authorization and only those authorized connections are allowed on a cluster. Authorization can be done by using the various privacy techniques such as IPSec, SSH, etc. [8].

The privacy on databases of big datasets are utmost important to protect and thus we can use the different privacy concepts such as k-anonymity, t-closeness, and l-diversity (these are the different models which protects the privacy) [8].

drksujatha23@gmail.com

To handle the privacy over big data many privacy-preserving techniques are available [9] Personal Information Identifier, Quasi-identifier, Sensitive Attribute, Non-sensitive Attribute Personal Information Identifier helps to identify the specific user [10]. Quasi-identifier means, in a database of the user if we use information of any user without losing confidentiality the different anonymous methods are used. First, the sensitive attributes such as social security number and number are removed from the database and then other attributes such as age, gender link with the external data which uniquely identifies the user and this attribute is called Quasi-identifier and to prevent the attacks in quasi-identifier we use k-anonymity technique [10].

3 Proposed Architecture

We discussed the following issues and challenges related to the big data security and privacy. By studying all the issues, we are now giving a new approach through which the privacy of the user's data may be protected after sending it to the Hadoop framework.

Privacy is important when more than one organization wants to perform some computation or analytics by anyone processing unit. Different clients are sending their data to the NameNode. After sending the data to the NameNode it gives it to Data Nodes for performing kind of processing on it. Here the Job Tracker and Task Tracker are used. Job tracker is responsible for sending the data to the Task Tracker. Task Tracker mainly works in a data node whose job is to process the job with the help of MapReduce function. The multiple clients are sending their data to the single NameNode so sometime NameNode performs malicious conduct and shares the data of one client to other for mutual benefit. Hence, the privacy may be violated.

So we proposed an architecture where different machines are to be present to perform analytics and the single client is sending data to different machines. Now, the client's data is distributed in a different machine and if malicious activity is done by computing machine then the machine is not able to provide whole of a client to others. As the data is available and distributed over the machines, so randomly we are choosing any NameNode for computation of result and the machine where the computation is to be performed takes the data from that machines and do analytics on that data. To perform a computation, randomized method is used to select one machine from the group.

By fetching the data from different machines for processing, no one knows to whom the data belongs. This happens because processing done by a machine which is randomly selected by the system and all other machines from the domain has sent all their data. The machine only knows the data as a whole for processing, not of an individual client for computation. Therefore, the identity of the user's data remains secure and due to which the privacy is to be maintained. In the proposed

drksujatha23@gmail.com

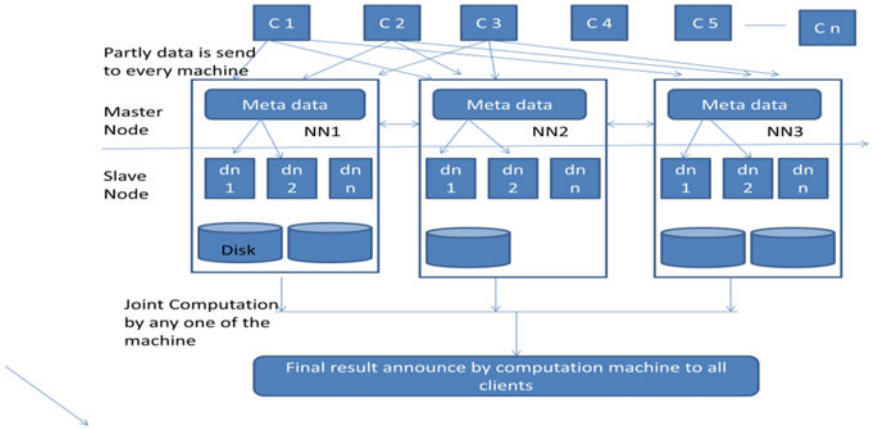


Fig. 1 Proposed architecture for big data privacy

architecture as shown in Fig. 1, Joint Computation is performed and that provides the privacy of the data from the computation machine or analytics machine.

We provide the result of security on the basis of the probability concept, where we can prove that as the number of machines for processing purposes increases the client's identity becomes more secure. If we have 1 machine then the probability is 1 that equals to 100% hence there will be 100% chance of machine to be vulnerable. We are using n no. of machines so the probability to disclose data of one client to others becomes $1/n$, which is very less. As machines for processing increases the probability of being vulnerable decreases deep down.

4 Conclusion

Big data is gaining popularity and advancement day-by-day but still, the security and privacy threats hinder the success of Big Data. It is necessary for every organization or every user to secure their data at every phase like generating, storing, and processing to keeps their data secure and protected. In this paper, we provide a survey on the certain types of issues and approaches related to the big data security and privacy. We also gave an idea which is useful for providing the privacy to the data after sending it to the analytics machine which provides computation and it somehow maintains the privacy of the user. The proposed architecture is very helpful in achieving the powerful privacy while performing big data analytics and during data processing data of different clients are private. In the future, we are going to simulate our architectural idea in different simulation machines to make it useful for real-time applications.

drksujatha23@gmail.com

References

1. Inukollu VN, Arsi S, Ravuri SR (2014) Security issues associated with big data in cloud computing. Trans IJNSA 6(3)
2. Savant VG (2015) Approaches to solve big data security issues and comparative study of cryptographic algorithms for data encryption. Trans IJICAR 1(1)
3. Terzi DS, Terzi R, Sagiroglu S (2015) A survey on security and privacy issues in big data. Trans ICITST 202–207
4. Cloud security alliance top ten big data security and privacy challenges. https://downloads.cloudsecurityalliance.org/initiatives/bdwg/Big_Data_Top_Ten_v1.pdf
5. How to manage big data's big security challenges. http://data-informed.com/manage-big-datas-big-security-challenges/
6. Mehmood A, Natgunanthan I, Xiang Y, Hua G, Guo S (2016) Protection of big data privacy. IEEE Access 4:1821–1834
7. Goyal V, Pandey O, Sahai A, Waters B (2006) Attribute based encryption for fine grained access control of encrypted data. In: Proceedings of ACM conference computer and communication security, Oct 2006, pp 89–98
8. Big data security and privacy handbook by cloud security alliance. https://downloads.cloudsecurityalliance.org/assets/research/big-data/BigData_Security_and_Privacy_Handbook.pdf
9. Mehta BB, Rao UP (2015) Privacy preserving unstructured big data analytics: issues and challenges, 11–12 Dec 2015
10. Search financial security. http://searchfinancialsecurity.techtarget.com/definition/personally-identifiable-information

drksujatha23@gmail.com

Stemmatizer—Stemmer-based Lemmatizer for Gujarati Text



Himadri Patel and Bankim Patel

Abstract Stemming is an important process in Information Retrieval (IR). Stem returned by stemmer need not be always a valid dictionary word. While a lemma returned by lemmatizer is always a valid dictionary word, which is a requirement of many IR systems. Indian languages are poor in resources. Specifically, the Gujarati language is having a stemmer but lacking a lemmatizer. In this paper, the authors have proposed 'The Stemmatizer'—stemmer-based lemmatizer for Gujarati language using a hybrid approach. It has the ability to learn new words. The proposed solution is tested on 2197 words and results have been found very much satisfactory.

1 Introduction

Manning et al. [7] proposed the basic steps for Information Retrieval (IR) in which morphological analysis is one of the important step, which includes stemming and lemmatization [6]. Stemming and lemmatization both aim to remove inflectional affixes and derivational suffixes. When the given word and stem/lemma belong to the same Part of Speech (PoS), it is called inflectional stemmer/lemmatizer. On the other hand, the derivational stemmer/lemmatizer term is used when they belong to different PoS [4, 12, 15]. However, the process of stemming and lemmatization is language dependent [12, 14]. Also, stemming is used as a substep of lemmatization when concern is to remove suffixes only. So, authors have used Stemmer to develop a lemmatizer.

Gujarati is Indo-Aryan language [3] having rich morphological features. It uses agglutination in its words [15]. For example, in word 'શિક્ષકની' (/ɕikʂəkəɳi/ - of

H. Patel (✉) · B. Patel
Shrimad Rajchandra Institute of Management and Computer Application,
Uka Tarsadia University, Bardoli, Surat, India
e-mail: himadri.patel@utu.ac.in; himadripatel87@gmail.com

B. Patel
e-mail: bankim_patel@srimca.edu.in

© Springer Nature Singapore Pte Ltd. 2019

V. S. Rathore et al. (eds.), *Emerging Trends in Expert Applications and Security*,
Advances in Intelligent Systems and Computing 841,
https://doi.org/10.1007/978-981-13-2285-3_78

drksujatha23@gmail.com

teacher), 'ની' is agglutinated with noun 'શિક્ષક'. Affix of a Gujarati word is identified based on its conceptual group or PoS [14, 15]. Gujarati words are in their either base form or inflectional form or derivational form. So, issues of lemmatizer need to be handled differently for it. Although Gujarati is a resource-poor language [8], several stemmers were developed so far [14]. The first Gujarati stemmer was developed using a hybrid approach [9] without considering derivational suffixes. Subsequently, stemmer has been developed by Suba et al. [15] with an improved result. A stemmer claiming the highest accuracy of 92.41% is Dhiya Stemmer [14] based on rule-based approach and tested using EMILLE corpus. However, it has the issue of overstemming and understemming. Also, loan words are not consistently stemmed. It is also noted in the literature [1] that overstemming is observed in 86% of total errors while 14% understemming is observed. So, comparatively addressing issue of overstemming will improve more in accuracy.

Lemmatizer developed for English language using Ripple Down Rule approach [11]. It is also developed for other European languages, but limited work found for Indian languages like Hindi, Punjabi, Oriya, Tamil, Bangla [10]. A rule-based Hindi lemmatizer [10] generate rules for extracting suffixes using knowledgebase by following the non-iterative method, which is unable to handle the words with multiple inflections. A similar method with the same issue is also used in [4] for developing Urdu lemmatizer. Multilingual lemmatizer [2] has been developed for 5 European languages and 18 Indian languages including Gujarati. They have used WordNet and tree-based data structure with backtracking provision carried out manually. It was tested for many Indian languages but excluded Gujarati. Except this, the authors have not come across any lemmatizer developed for the Gujarati language. So, a lemmatizer is required in Gujarati which also addresses the issues of different lemmatizer discussed above.

Study of state of the art shows that rule-based approach is used in the Stemmer [9, 14, 15] and lemmatizer [4, 5, 10, 11] developed in Indian languages. Also, lemmatizer [4, 5, 10] developed in Indian languages use a dictionary of words. Referring to this, authors are using rule-based approach with a combination of dictionary-based approach.

2 Issues in Lemmatization of Gujarati Language Text

A Gujarati word is a lemma suffixed with several inflections like ની, ઓ, માં, etc. Although Gujarati follows SOV typology [3], it is free order [8] and highly inflected [14], morphologically rich [15] language. Having these characteristics, finding lemma of a Gujarati word by removing inflections is very much benefitted in information retrieval [7]. Issues related to lemmatization of Gujarati language text are discussed below:

i. Gujarati word has more than one inflection [13]. So, usual single iteration is not sufficient to identify all inflections used in the word.

drksujatha23@gmail.com

ii. A lemmatized word derived after removing inflection may not be a lemma but only a stem. A word 'શાળામાં' (/ɛɑlɑmɑ̃/ - in the school) gives 'શાળ' (/ɛɑlə/) which is not a lemma. The lemma is 'શાળા' (/ɛɑlɑ/ - school). Such cases need to be carefully handled.
iii. Only removing inflection does not give a root word always. A word 'વ્યાવસાયિક (/ʋjɑʋəsɑ̪jik/ - professional) gives incorrect word 'વ્યાવસાય' (/ʋjɑʋəsɑ̪jə/), for the correct root word 'વ્યવસાય' (/ʋjəʋəsɑ̪jə/ - profession). These words are derivationally inflected having infixes. These infixes need to be identified and processed appropriately.
iv. Some inflections are themselves lemmas. The inflection 'હાર' (/ɦɑrə/ - defeat) is a lemma, which appears in words like 'તારણહાર' (/ʈɑrəɳəɦɑrə/ - savior). So, if a word to be lemmatized is itself an inflection, it removes the whole word.
v. New word received after removing inflection from the given word may be a lemma having a different meaning. A word 'સહેલો' (/sɑ̪ɦeːloː/ - easy) gives a lemma 'સહેલ' (/sɑ̪ɦeːlə/ - trip), while the correct lemma for the word is 'સહેલું' (/sɑ̪ɦeːlu̯ɑ̃/ - easy).

Considering the above issues, existing approaches to find lemmatizer discussed above may not work for Gujarati. These issues are addressed in this paper and accordingly, Stemmatizer is developed using a hybrid approach by combining rule-based with dictionary-based approach with the scope of continuous learning of new words and tested on 2197 Gujarati words from different domains.

3 Proposed Methodology

The Stemmatizer developed in this paper uses a vocabulary to derive a lemma, which should be a linguistically meaningful dictionary word. Here, a valid dictionary word is valid lexically as well as semantically. In addition, the proposed Stemmatizer is both inflectional and derivational.

3.1 Description of Vocabulary

Authors have created a vocabulary which plays important role in resolving the above discussed issues. It stores both stem and lemma for a word as the proposed algorithm uses the stem to acquire a lemma. The basic structure followed in vocabulary is <stem>$<lemma>, where $ is used as a separator. The vocabulary is structured as shown in Fig. 1 and a description is given below:

i. Stem-lemma combination is stored in the above-specified format for the words whose stem and lemma are same. For example, 'શિક્ષક' (/ɛikʂəkə/ - teacher) is a lemma which is also a stem of words 'શિક્ષકો' (/ɛikʂəkoː/ - teachers), 'શિક્ષકની' (/ɛikʂəkənɨ̃/ - of teacher), 'શિક્ષકોને' (/ɛikʂəkoːɳɛː/ - to teachers) etc. So, in this case,

drksujatha23@gmail.com

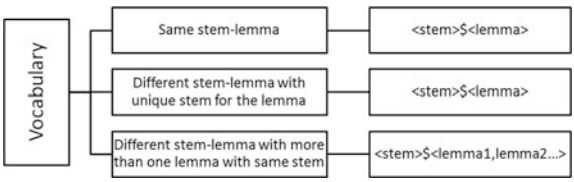


Fig. 1 Structure of vocabulary

vocabulary only stores 'શિક્ષક$શિક્ષક'. Also, the authors store a combination of stem-lemma for the words whose stem and lemma are different but there is no other word with different meaning whose stem is the same as of this word. For example, 'છોકર' (/ʧʰo:kərə/) is a stem and 'છોકરું' (/ʧʰo:kərũ/ - kid) is a lemma of words 'છોકરાઓ' (/ʧʰo:kərɑo:/ - boys), 'છોકરી' (/ʧʰo:kəri/ - girl), 'છોકરુંને' (/ʧʰo:kərũɳɛ:/ - to kid) etc. So, vocabulary stores 'છોકર$છોકરું'. If the Stemmatizer is given a word 'શાળામાં' (/ɛɑ|ɑmɑ̃/ - in the school) it gives 'શાળ' (/ɛɑ|ə/), which is a stem of the word. This stem is matched with the record 'શાળ$શાળા' stored in vocabulary and returns its correct lemma 'શાળા' resolving the second issue discussed above. Similarly, vocabulary stores 'વ્યાવસાય$વ્યવસાય' to resolve third issue discussed above.

ii. Cases where more than one lemma with different meaning having the same stem, need to be handled differently. In such cases, stem is not unique, so all possible lemma for that common stem are stored in little modified format as <stem>$<lemma1>, <lemma2>. For example, 'રાજી' (/rɑdʒi/ - cheerful) and 'રાજ' (/rɑdʒə/ - governance) are lemma with different meaning giving same stem 'રાજ' (/rɑdʒə/). So, vocabulary stores 'રાજ$રાજ,રાજી'. When a word 'રાજી' given to Stemmatizer, it matches with the record 'રાજ$રાજ,રાજી'. In this case, the Stemmatizer returns both the lemmas ensuring to address ambiguity in the lemma of the inputted word. This way the fifth issue discussed above is resolved.

Both structures of vocabulary are maintained separately to optimize the algorithm discussed below by reducing steps. Also, it is in ascending order of stem to make searching faster.

3.2 Algorithm

Totally, 179 inflections [13] are checked either as a rule of removal or as a rule of replacement [14]. The list of inflections also contains 52 inflections used in [14]. Following rules are used for implementing the Stemmatizer:

i. Both structures of vocabulary are searched either for a match with the given word to check if it is itself a lemma or with a list of inflections.

drksujatha23@gmail.com

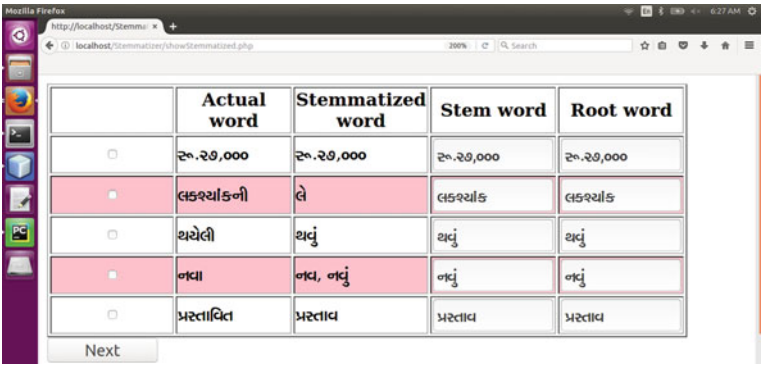


Fig. 2 List of stemmatized words

ii. All inflections are matched recursively to search for any inflection in any order until no inflection matches with the word. When any inflection is matched, it is removed from the word. The recursive searching ensures removal of all inflections resolving first issue.
iii. The word fully matched with any of the inflection is not removed to handle fourth issue.
iv. Only rule-based approach is used for the word not found in vocabulary.

The implementation of Stemmatizer is done in Python 3.x with PHP 5.x. It accepts a filename as input and shows all stemmatized words as an output as shown in Fig. 2. It also allows selecting the wrongly stemmatized words using a checkbox as shown in Fig. 2, editing them to make them correct and adding them to a vocabulary at the appropriate position. In Fig. 2, 'Actual word' is an inputted word, 'Stemmatized word' is the result given by Stemmatizer, 'Stem word' and 'Root word' fields are textboxes which allows correcting the wrong stem and lemma, then storing it to vocabulary. The concept of learning new words in vocabulary is described in the next section. In Fig. 2, the fourth word 'નવા' is stemmed as 'નવ' which is matched with entry in vocabulary 'નવ$નવ, નવું'. So, it is returning two words as stemmatized words. Other words are giving the single correct word as stemmatized word.

3.3 *Learning of New Vocabulary*

Being a resource-poor language [8], Gujarati is lacking the availability of a dictionary in a specific format to use in this research. So, authors have developed a vocabulary manually, which may not include all words. As a solution, Stemmatizer provides an option to store the missing word in the vocabulary. Before storing, it

drksujatha23@gmail.com

allows to edit the stem and lemma of the word if it is returned incorrect by rule-based approach. The new word is placed at an appropriate position in vocabulary so that it gives correct output when appeared next time. In Fig. 2, the second word 'લાફ્યાક' is wrongly spelled the word for the correct word 'લાફ્યાક'. So, it is not matched with the vocabulary and giving incorrect output. The Stemmatizer allows editing by checking the checkbox and inputting the correct stem and lemma. Only words with checked checkboxes are added to vocabulary.

4 Result Analysis

Totally, 2197 words are considered for testing the stemmatizer. These words are from domains education, politics, religion, and business. Out of which there are 46 words having a stem with two or more lemma in the dictionary, 8 combined words using '–' or '.' and 46 wrongly spelled words. So, total 2097 words were tested on the Stemmatizer, out of which 35 words were incorrectly identified. This gives an accuracy of 98.33%. While testing the stemmatizer, 239 words were not found in the dictionary, so they were added to the vocabulary using its learning ability. Table 1 shows the result analysis of the tested words by bifurcating them based on words found in vocabulary and not found in vocabulary. The highest accuracy is achieved when words are found in vocabulary.

Table 2 shows the result analysis of tested words by bifurcating them according to their domains. In this analysis, words may belong to more than one domain. According to the analysis, the words of 'Politics' domain is giving highest accuracy.

Table 1 Result analysis (vocabulary-status wise)

	Categorywise total words	Correctly identified	Wrongly identified	Accuracy (%)
Found words in vocabulary	1821	1811	10	99.45
Not found words in vocabulary	276	251	25	90.94
Total tested words	2097	2062	35	98.33

Table 2 Result analysis (domain wise)

Domain	Categorywise total words	Correctly identified	Wrongly identified	Accuracy (%)
Education	563	551	12	97.87
Politics	586	578	8	98.63
Business	560	550	10	98.21

drksujatha23@gmail.com

The wrong result given by the Stemmatizer is mainly due to two reasons: (i) for some words, its inflected word or part of the inflected word is stored in the dictionary as another lemma. For example, 'જ્યાં' (/dʒe:mɑ̃/ - in which) is having stem 'જે' with inflection 'યાં'. But after removing inflections 'ો+ં', it gives 'જેમ' (/dʒe:mə/ - like) which is also a stem having a different meaning. In this case, it returns 'જેમ' (/dʒe:mə/ - like) instead of 'જે' (/dʒe:/ - which). (ii) For some inflections, the part of it appears as another inflection which is tested before the said inflection. For example, word 'બળપૂર્વક' (/bəl̪əpurʊəkə/ - forcefully) is having inflection 'પૂર્વક' (/purʊəkə/). But due to the sorted order of inflection according to its length, another inflection 'ક' is found and removed. So, actual inflection cannot be found and removed. This gives output as 'બળપૂર્વ' (/bəl̪əpurʊə/) which is incorrect.

5 Conclusion and Future Enhancement

Authors have developed Stemmatizer—a lemmatizer based on stemmer using rule-based approach with a combination of dictionary-based approach. It is using a vocabulary having predefined format. The stemmatizer is capable to learn new words. It is having an accuracy of 98.33%. The points discussed in above section of result analysis for the incorrect output of the Stemmatizer are still to be improved. Also, the algorithm of the Stemmatizer is not optimized with respect to time.

References

1. Ameta J, Joshi N, Mathur I (2012) A lightweight stemmer for Gujarati. In: Proceedings of 46th annual convention of computer society of India
2. Bhattacharyya P, Bahuguna A, Talukdar L (2014) Facilitating multi-lingual sense annotation: human mediated lemmatizer. In: Global WordNet conference
3. Deo A, Sharma D (2006) Typological variation in the ergative morphology of Indo-Aryan languages
4. Gupta V, Joshi N, Mathur I (2015) Design and development of rule based inflectional and derivational Urdu stemmer 'Usal'. In: 2015 International conference on futuristic trends on computational analysis and knowledge management (ABLAZE). IEEE, pp 7–12
5. Gupta V, Joshi N, Mathur I (2016) Design and development of a rule-based Urdu lemmatizer. In: Proceedings of international conference on ICT for sustainable development. Springer, Singapore
6. Jurafsky D (2000) Speech and language processing. In: Daniel J (ed) Speech and language processing. Pearson Education India
7. Manning C, Raghavan P, Schütze H (2008) An introduction to information retrieval. Cambridge University Press, Cambridge
8. Patel H, Patel B (2016) A critical study of challenges in educational opinion mining of text written in Gujarati language. Natl J Syst Inf Technol 25–34
9. Patel P, Popat K, Bhattacharyya P (2010) Hybrid stemmer for Gujarati. In: 23rd International conference on computational linguistics

drksujatha23@gmail.com

10. Paul S, Tandon M, Joshi N, Mathur I (2013) Design of a rule based Hindi lemmatizer. In: Proceedings of third international workshop on artificial intelligence, soft computing and applications. Chennai
11. Plisson J, Lavrac N, Mladenic D (2004) A rule based approach to word lemmatization
12. Prathibha RJ, Padma MC (2015) Design of rule based lemmatizer for Kannada inflectional words. In: International conference on emerging research in electronics, computer science and technology, pp 264–269
13. Rathod B, Shah P (2017) Gujarati Vyakran Parichay. Akshar Publication, Ahmedabad
14. Sheth J, Patel B (2014) Dhiya: a stemmer for morphological level analysis of Gujarati language. In: 2014 International conference on issues and challenges in intelligent computing techniques (ICICT). IEEE
15. Suba K, Jiandani D, Bhattacharyya P (2011) Hybrid inflectional stemmer and rule-based derivational stemmer for Gujarati. In: Proceedings of the 2nd workshop on south and southeast Asian natural language processing

drksujatha23@gmail.com

Performance Impact on Different Parameters by the Continuous Evolution of Distributed Algorithms in Wireless Sensor Networks: A Study



Hemlata Soni, Gaurav Gupta and V. K. Chandna

Abstract Wireless sensing network is one among the rising networks that work underneath inherent resource constraints. Topologies of those forms of networks will rework dynamically supporting the placement, variety of sensing element nodes, and application. It is indispensable to create effective distributed algorithms to handle the energy, information measure limitations of WSNs. With time, variety of algorithms have evolved for WSNs. However, it is exhausting to mention that anyone of the prevailing algorithms is optimum for these networks in each and every situation. We have a tendency to study many algorithms for WSNs that majorly influenced the performance of these networks. For each new algorithm built, one or additional Quality of Service (QoS) parameter of the network proportionally improved the performance of the network. This paper emphasizes on providing a summary of the schemes proposed by different scholars for the development of energy-efficient distributed algorithms and improved QoS parameters of the network at one place.

1 Introduction

WSN could be a quickly growing field that integrates sensing, computation, and communication into one small device. At constant time, the capabilities of a selected single device is minimum, the composition of many devices offers essentially new technological prospects [1]. The ability of WSNs lies within the capability organize and to prepare an immense quantity of little nodes that garner and arrange themselves. These devices square measure exploited in many

H. Soni · G. Gupta (✉) · V. K. Chandna
Jaipur Engineering College and Research Centre, Jaipur, India
e-mail: gauravgupta.cse19@jecrc.ac.in

H. Soni
e-mail: hemlata.soni.1982@ieee.org

V. K. Chandna
e-mail: vinaychandna@yahoo.co.in

© Springer Nature Singapore Pte Ltd. 2019

V. S. Rathore et al. (eds.), *Emerging Trends in Expert Applications and Security*,
Advances in Intelligent Systems and Computing 841,
https://doi.org/10.1007/978-981-13-2285-3_79

drksujatha23@gmail.com

applications that vary from period of time observance to observation of environmental conditions, present computing atmosphere. At constant time, WSNs are usually recognized as management actuators that widen management from the Net into the physical world [2].

At present, WSNs show the proof of an implausible analysis momentum. Scientists and research engineers from the majority of the fields grasp this explicit field. It is concerned by researchers working in the field of hardware technology, operational systems, middleware, graph theory, and pure mathematics. Several fundamental scaling laws, say, the ability of the device network is traced by data and communication researchers. Currently, database and network researchers are developing typical databases and new protocols, respectively, for sensor networks.

As a sensor network represents a message passing graph and it has been the major focus. Thus, it is believed that sensor networks will adapt distributed algorithms directly or at any pace [3, 4].

Recent past shows that distributed algorithms are associated with nursing increasing analysis field. Sensor Networks are less at risk of facet effects, which make it distinct from various other natural application areas such as peer-to-peer networks, Internet, etc. As an example, the narcissistic behavior of individual device nodes, unremarkably the whole network is in hand by one entity [5].

Within the computer networks and structural design and operation of knowledge, MST problem often occurs. The tree with the lowest weight among all possible spanning where the sum of the weights of the edges in the tree accounts to the weight of the spanning tree is known as Minimum Spanning Tree. For data aggregation, MST is the mostly preferred routing tree in an ad hoc sensor network [6]. In distributed algorithms, the amount messages being communicated among the nodes and the running time determines its effectiveness and design of optimal algorithms with respect to these conditions have been considered by several researchers. Square measures distributed algorithms that discover the local time [7, 8] and square measures are basically optimal and they supported the time quality.

In many applications, the ad hoc fashion helps to form the network by native self-configuration as within the region of interest the sensors are scattered liberally. Energy is especially an elementary resource owing to battery restrictions. Additionally, the communication and network management should be disbursed in a distributed and native fashion because each sensor recognizes its neighbors. Distributed algorithms are not applicable in an energy-constrained sensor network. Once the network needs to be reassembled recurrently and quickly, this is often notably true in a very dynamic setting as interchanging huge number of messages comparatively utilizes a large amount of time and energy. To extend the network lifetime, systematic distribution of energy among nodes is essential and thus leading to reconfiguration [6].

drksujatha23@gmail.com

2 Proposed Minimum Spanning Tree Algorithms

2.1 Choi et al. [9]

The paper presented by Choi [11] addressed an important distributed computing problem of the minimum spanning tree. He studied the energy-efficient distributed algorithms, which showed a nontrivial lower bound of **Ω log n** on the energy-complexity by assuming a random distribution of nodes. The above results are an advancement of average energy complexities of previously best-known algorithms (Fig. 1).

2.2 Khan et al. [10]

As several distributed algorithms needed a larger number of messages and time, Khan et al. [12] in wireless ad hoc network for construction of a minimum spanning trees investigated and designed a simple and local class of energy-efficient algorithms called Nearest Neighbor Tree (NNT) algorithms. The algorithm gave close resemblance to the MST. The author showed that with a poly-logarithmic number of rearrangements (node insertions/deletions), the NNT Algorithm can be continued dynamically.

2.3 Wang [11]

Wang [13] improved the approach of getting an MST from the entire spanning tree by presenting a new algorithm which used the characteristics and binary code of MST. Initially, this approach removed some of the non-spanning trees based on the number of the graph edges. Furthermore, it eradicates some of the non-spanning trees based on the judgment of the graph connectivity. In reality, the fundamental

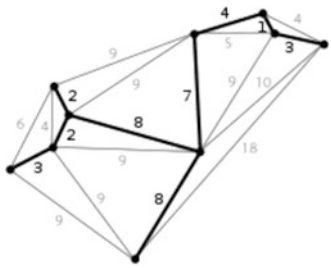


Fig. 1 Visualizing minimum spanning tree

drksujatha23@gmail.com

work of this algorithm is to search for the most excellent in the global scope. Finally, it is simple to discover all the MSTs of the connected graph considering the algorithm.

2.4 Sanger and Agarwal [14]

In the operation of network design, the MST problem is a standard problem. Its bi-objective versions are non-deterministic polynomial-time hard and thus could not be solved efficiently. Sanger and Agarwal [14] compared three tree encoding approaches based on a bi-objective evolutionary algorithm. To provide a solution for three different instances of bi-objective MST problem, evolutionary approaches in tree encoding techniques are utilized. Thus, on the account of obtained Pareto optimal front, tree encoding techniques study is carried out accordingly. The approach embeds the bi-objective MST problem by means of Non-Dominated Sorting Genetic Algorithm II (NSGAII) (Fig. 2).

2.5 Wang et al. [13]

In practice, clustering algorithms based upon minimum spanning tree are used extensively because of its potential to detect clusters with unbalanced boundaries. The chief source of computation in clustering algorithms is nearest neighbor search and the standard solutions take $O(N^2)$ time. Wang et al. [15] studied them and developed a fast minimum spanning tree-inspired clustering algorithm which utilized cut property and effectively implemented cycles of MST thus performing better than the $O(N^2)$.

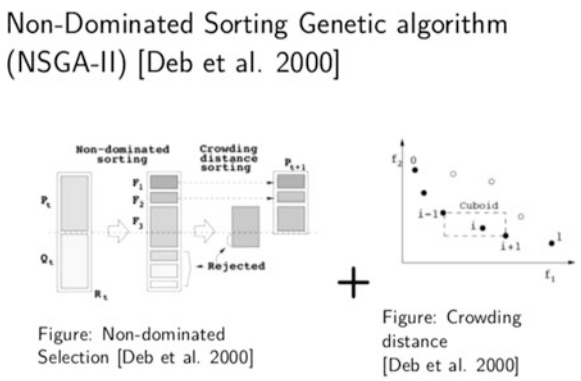


Fig. 2 Nondominating sorting genetic algorithm

drksujatha23@gmail.com

2.6 Gong and Jiang [14]

Gong and Jiang [14] proposed an algorithm intended to extend the lifetime of WSNs by efficiently exploiting the energy resources of sensor nodes present in inadequate amount. The tree routing protocol called TRP, based on power-aware trees calculated the distance between neighboring nodes and sink. The algorithm leads to build the best possible MST rooted on the sink (Fig. 3).

2.7 Ghosh et al. [15]

Ghosh et al. [15] investigated TDMA-based sensor networks by measuring increment in the aggregated data collection rate and reduction in the maximum packet delay and thus realizing the consequence of routing topologies. The author confirmed that to accomplish best comparison between packet delays and data collection, trees with minimum radius and bounded node degree were required. Thus, to obtain a solution of best possible routing tree construction problem, they developed a bi-criteria formulation.

2.8 Yanrong Cui and Hang Qin [16]

Base station exchanges information with every node that communicates to it in various WSN applications. Whenever a node exchanges information with base station directly, nodes' energy is wasted as there could be various duplications of data and information. In order to overcome this issue, Cui and Qin [16] put forward Data Query Protocol based on Minimum Spanning Tree (QPMST). This protocol considerably reduces the transmission of redundancy data, saves the nodes' energy and increases the network lifetime. It constructs minimum spanning tree where the root is the sink node and children as supplementary nodes. In the protocol,

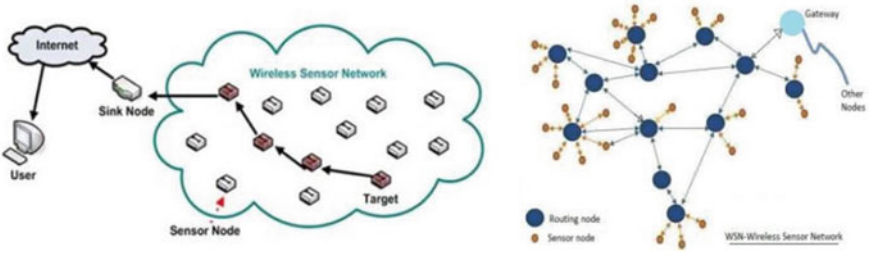


Fig. 3 Diagrams depicting sensor node and sink nodes

drksujatha23@gmail.com

Table 1 An overview of the latest work in the distributed algorithm for MST

Author	Parameters used for improving performance
Zhang et al.	The proposed algorithm, which finds disjoint paths between two nodes, namely Distributed Finding Disjoint Paths (DFDP)
Khilar and Meenakshi Panda	The proposed algorithm searched WSNs for the presence of hard and soft faulty sensor nodes
Rosana et al.	The work evaluated and implemented an algorithm for fault tolerance and load balancing using multiple routes between sensor nodes by constructing routing trees
Zhang	The paper proposed a way to examine topological information by using advanced distributed layered negotiation based approximation algorithm
Minlan et al.	They worked upon multidimensional scaling of anchor nodes and developed a 3D WSNs node localization algorithm. The distance between the anchor and other network nodes is calculated using the algorithm
Zhang et al.	The paper gave a fuzzy approach based distributed clustering algorithm which was energy efficient having nonuniform distribution
Khan et al.	The authors proposed an energy-efficient WSN protocol consisting of a routing algorithm for the transmission of data, cluster head selection algorithm, and a scheme for the formation of clusters

sink node broadcasts query tasks and the results are taken by sink from leaf nodes and stored in their parent node (Table 1).

3 Conclusion

This paper discusses different algorithms of WSNs which evolved with time for better and efficient performance of the network. Different algorithms improved performance by variations of different Quality of Service parameters such as load balancing, fault tolerance, cluster formation, etc. Thus the paper assists the future researches by giving a brief view of energy-efficient distributed algorithms developed in the recent decade and the major changes in parameter that optimized the performance of the Wireless Sensor Networks.

References

1. Hill JL (2003) System architecture for WSNs, Dissertation. University of California, Berkeley
2. Shankar M, Sridar M, Rajani M (2012) Wireless sensor network safety study. Int J Comput Technol Appl 3(1):160–176

drksujatha23@gmail.com

3. Uddin M, Shah A, Alsaqour Raed, Memon J (2013) Measuring efficiency of tier level data centers to implement green energy efficient data centers. Middle-East J Sci Res 15(2):200–207
4. Tabrizi HB, Abbasi A, Sarvestani HJ (2013) Comparing the static and dynamic balances and their relationship with the anthropometrical characteristics in the athletes of selected sports. Middle-East J Sci Res 15(2):216–221
5. Schmid S, Wattenhofer R (2006) Algorithmic model for sensor networks. In: 14th international workshop on parallel and distributed real-time systems (WPDRTS). Island of Rhodes, Greece
6. Krishnamachari B, Estrin D, Wicker S (2002) The impact of data aggregation in wireless sensor networks. In: Proceedings of the second international workshop distributed event-based system (DEBS 02)
7. Elkin M (2004) Unconditional lower bounds on the time-approximation tradeoffs for the distributed minimum spanning free problem. In: Proceedings of the 36th ACM symposium on theory of computing STOC'04
8. Peleg D (2000) Distributed computing: a locality sensitive for industrial and applied mathematics
9. Choi Y, Pandurangan G, Khan M, Kumar VSA (2009) Energy-optimal distributed algorithms for minimum spanning trees. IEEE J Sel Areas Commun 27:1297–1304
10. Khan M, Pandurangan G, Anil Kumar VS (2009) Distributed algorithms for constructing approximate minimum spanning trees in wireless sensor networks. IEEE Trans Parallel Distrib Syst 20:124–139
11. Wang F (2011) A minimum spanning tree algorithm based on binary coding. In: International conference on multimedia technology (ICMT), pp 5227–5229
12. Sanger AKS, Agarwal AK (2010) Comparison of tree encoding schemes for bi-objective minimum spanning tree problem. In: 2nd IEEE International conference on information and financial engineering (ICIFE), pp 233–236
13. Wang X, Wang X, Wilkes DM (2009) A divide-and-conquer approach for minimum spanning free-based clustering. IEEE Trans knowl Data Eng 21:945–958
14. Gong B, Tingyao J (2011) A tree-based routing protocol in wireless sensor networks. In: International conference on electrical and control engineering (ICECE), pp 5729–5732
15. Ghosh A, Incel OD, Kumar VSA, Krishnamachari B (2010) Bounded-degree minimum-radius spanning trees for fast data collection in sensor networks. In: IN-EOCOM IEEE conference on computer communications workshops, pp 1–2
16. Cui Y, Qin H (2010) Data query protocol based on minimum spanning tree for wireless sensor network. In: Fourth international conference on genetic and evolutionary computing (ICGEC), pp 798–801

drksujatha23@gmail.com

A Framework of Lean ERP Focusing MSMEs for Sales Management



Shilpa Vijaivargia and Hemant Kumar Garg

Abstract This paper presents a framework of lean ERP systems focusing on MSMEs. Most of the MSMEs do not have requisite infrastructure and finance to adopt ERP systems available in the market or get customized ERP systems as per their requirements. Even if procured, it is difficult to maintain such ERP due to lack of in-house IT skilled manpower and finance for its maintenance. The concept of lean ERP is supported by individual module for each core and support function performed by MSMEs. These modules have non-rigid interlinking with each other supporting Service-Oriented Architecture (SOA) and form an ERP system by combining together. Lean ERP concept, its generic framework, ERP modules and artefacts of one of its modules sales management were explained through workflows, norms, routines, Standard Operating Procedures (SOPs) and database configuration in this paper.

Keywords MSME · Small industries · Lean ERP · Sales management Artefacts

1 Introduction

MSMEs play a vital role in Indian economy; the estimated contribution of Micro, Small and Medium Enterprises (MSMEs) to India's Gross Domestic Product (GDP) including manufacturing and service sector is on increasing trend [1]. It is estimated 37.54% in FY 2012–13, 42.38% in FY 2013–14 and 44.70% in FY 2014–15 [1]. Total employment in the MSME sector is more than 805.24 lakhs [2].

S. Vijaivargia (✉)
Department of Scientific and Industrial Research, Ministry of Science and Technology, New Delhi, India
e-mail: shilpavijay18@gmail.com

H. K. Garg
Government Women Polytechnic College, Jaipur, Rajasthan, India
e-mail: hgarg1710@gmail.com

© Springer Nature Singapore Pte Ltd. 2019

V. S. Rathore et al. (eds.), *Emerging Trends in Expert Applications and Security*, Advances in Intelligent Systems and Computing 841,
https://doi.org/10.1007/978-981-13-2285-3_80

drksujatha23@gmail.com

Even with large investment and focus of the Government on skilling up MSMEs, it has been noticed that these sectors lack in producing deliverables as desired. One of the major reasons behind it is lack of appropriate mapping of manpower, skills and processes in their competence metric which would result in delay in producing deliverables as benchmarked in terms of quality and quantity both. Most of the MSMEs in India are unaware about digitalization of processes, cost management and skill mapping and presently handling their core and support functions traditionally. Important milestones in production of goods and services by MSMEs such as inventory management, sales and marketing, production, procurement and stores do not follow any standard procedures which would result in lack of efficiency, speed and transparency in desired outcomes [3]. Due to these reasons, the growth rate of MSMEs has not been achieved as desired.

Hence, a strategic approach and IT-based ERP model need to be formulated which would help MSMEs in set of performance standards over various aspects of production, sales and services in order to optimize the use of resources, achieve better quality, shorten working capital cycles, discreet purchase of raw materials, leverage market intelligence and maximize their returns with effective product delivery to reach marketplace.

Lean ERP system with minimal workflows suggested as a solution for this sector providing capability for data input, process and information to meet the requirements of employees and other stakeholders on various roles and responsibilities.

Lean Concept

Lean in the context of MSMEs presents the convergence of a number of manufacturing philosophies such as just-in-time, quality at the source and total quality management. Lean manufacturing uses less human effort in the factory, manufacturing space, capital investment and time interval between customer order and shipment [4].

The concept of lean ERP presented in this paper is based on study of common procedures followed by MSMEs considering *Central Electronics Limited (CEL)* as an illustrative study model. In Sect. 2, background and related work are presented. In Sect. 3, framework for lean ERP is presented followed by artefacts for *Sales Management* module in Sect. 4, research outcome is given in Sects. 5 and 6 presents conclusion and future work.

2 Background and Related Work

As is Study

As is study was done for Central Electronics Limited (CEL), a Government of India Enterprise to understand business processes. Organizational processes were mapped and re-engineered to support the concept of lean ERP. Business processes and modules were developed based on input forms, physical observation and

drksujatha23@gmail.com

interaction, personal interview and secondary research with an objective to prepare Standard Operations and Procedures (SOPs) under lean ERP concept. Time spent in accomplishing a process was also helpful in formulating SOPs to automate the present operating procedure.

To-be State

The to-be state presents optimization of processes considering re-engineering, solution development and the possible effects of information technology use on lean times and operational costs.

3 Framework for Lean ERP

ERP application is vital and complex in nature; the goal was to keep lean ERP concise with minimal important modules to make work them as standard modules. Generic workflows were created for *Sales Management* module which could be implemented in MSMEs. The concept supports Service-Oriented Architecture (SOA) in which one or more standard modules may be taken by MSMEs on subscription basis as per their requirements.

Lean ERP Prototype

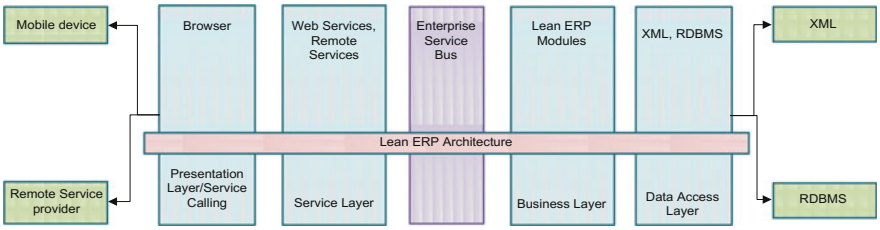
Operation metrics, standards, parameter classification and functions used in the organization were taken into consideration while designing the artefacts of *Sales Management* module. Data fields were incorporated within the input forms to capture the desired and actual output of the different resources. The fields/attributes incorporated with them were reviewed with the existing available input forms in order to verify the correlation between them. The comparison of actual to desired value is used to set high targets pertaining to getting orders and reduced cycle time for each activity.

Application Architecture

SOA design principles were used in lean ERP concept. The objective is to make business processes more flexible, explicit service, data and application integration. The database will consist of metadata defined by a set of business rules, SOPs and logic flows. Functions for common services were kept as by their name such as Get_Customers, Get_orders and Create_Invoice, which were being part of the *Sales Management* module (Schematic 1).

In the above architectural approach, the interoperable business services are not coupled rigidly to facilitate them to share within enterprises. It provides flexibility in building or upgrade existing services. A single version of the application with a

drksujatha23@gmail.com



Schematic 1 Lean ERP architecture

single configuration (hardware, network and operating system) can be used for all customers as the application supports installation on multiple machines (horizontal scaling) to support scalability.

4 Artefact for Sales Management Module

The module *Sales Management* describes the sequence of process steps or activities related to sales and inventory management. The workflow defined is given (Schematic 2).

The request for material is placed by the marketing department by filling material issue requisition form. The stock gets checked by ERP and an alert is sent to stock manager to transfer the required material to despatch section. The despatch section supplies the required material to customer and an alert gets generated to customer. Based on feedback received from the customer, if accepted the stock gets updated through the system else, if required, the material gets replaced from the stores.

Estimated Results Against the Applied Tool

The tool was applied to raw material related to solar photovoltaic where stock varied from 400 to 1,500 averages per week.

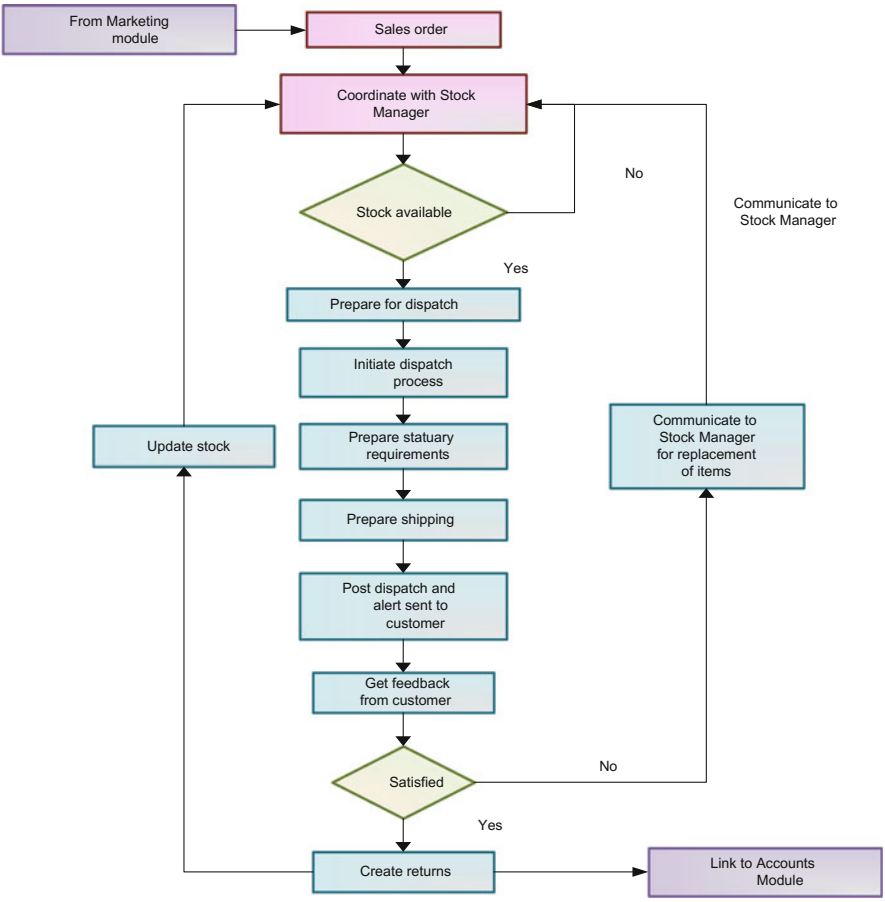
1. Reorder level is the time which indicates when order for raw material is required which could be in between of maximum and minimum inventory level.

   Reorder level (units) = Maximum Lead time (in weeks)* Maximum Usage (Max units used per week)

$$7,000 = 10 * 700 \text{ where; } \begin{cases} \text{Maximum Lead time} = 10 \text{ weeks} \\ \text{Maximum usage} = 700 \text{ units per week} \end{cases}$$

2. Optimum quantity that should be bought considering all the other cost factors for a particular product.

drksujatha23@gmail.com



Schematic 2 Sales order processing

Reorder quantity (Units) = EOQ (Economic Order Quantity) = sqrt {(2 * A * O)/C}

Sqrt(2 * 15,000 * 1,20,000/15,000)

= 489.9 ∼ 490

where
A is estimated annual requirement = 15, 000
O is estimated ordering cost per year = 1, 20, 000
C is estimated carrying cost per material per year = 15, 000

drksujatha23@gmail.com

5 Research Outcome

Lean ERP framework focusing on MSMEs was built for *Sales Management* module, which is one of the modules of lean ERP defined in this paper. Workflow related to *Sales Management* module was presented considering operation metrics, SOPs, inventory and process cycle. Activity workflows, norms, database configuration and architecture of *Sales Management* module were presented in the paper. Results were calculated taking estimated inputs values for raw material for optimum quantity of raw material to be bought. Process workflows for *Sales Management* module follow the practice of continuous flow production, minimum inventory, quality of deliverables and auto updates such as inventory management, which provide the key functions necessary to update and maintain raw materials. These allocations are relieved as inventory items get issued to the jobs or purchase order receipts are posted.

6 Conclusion and Future Work

Implementation of lean ERP would help MSMEs to improve their business metrics by process optimization, improving supply chain process, integration across various functionalities and increasing productivity inculcating a culture of continuous improvement. It will also support transparency in execution of processes through digital mode as per the need of the hour and reduce the cost incurred in comparison to execution of manual processes. Such solution may be deployed on cloud environment as the concept is also supported by Ministry of MSME through its various schemes for MSMEs such as scheme on support for lean manufacturing, scheme on support for deployment of ERP on cloud environment on subsidized rates, scheme for 'Technology and Quality Up-gradation (TEQUP) Support to MSMEs', etc. [2].

Adopting a lean manufacturing/lean production methodology would help MSMEs aims at continuous elimination of waste in production process, lowering overall production costs; increased productivity output; and shorten production lead time. The above illustrative study target may act as a model for MSME units, which may implement this model to improve and increase the productivity and efficiency of their processes and resources. As similar to *Sales Management* module, the architecture of other modules on lean ERP concept may be developed and implemented for MSMEs as future work.

drksujatha23@gmail.com

References

1. Press Information Bureau, Government of India. http://pib.nic.in/newsite/PrintRelease.aspx?relid=107201
2. www.dcmsme.gov.in
3. Building capabilities for implementation of lean ERP/Automated solutions in PSUs/MSMEs, CDC project report by M/s vayamtech
4. LeanERP® mobile platform solution for planning, visualization and execution of business operations in MSME units. Int J Electron Commun Comput Eng 4(1). ISSN (Online): 2249–071X, ISSN (Print): 2278–4209

drksujatha23@gmail.com

Multimedia Cloud for Higher Education Establishments: A Reflection



Anjum Zameer Bhat, Vikas Rao Naidu and Baldev Singh

Abstract Cloud computing has been a revolutionary technology in the modern era of information and communications technology. It has in the recent past created a significant impact on how computing resources can be utilized without necessarily investing a huge amount of money and without any requirements of administrative overhead that is usually the case with in-house infrastructure building. It has revolutionized the way of IT infrastructure development, implementation and deployment. Multimedia is one of the key specialization areas in computer graphics, which is taught in most of the professional engineering colleges across sultanate of Oman. The multimedia degree and diploma courses are mostly based on applied multimedia, and henceforth, many multimedia applications are taught throughout the tenure of the bachelor's or diploma course. Most of the multimedia applications are very costly and a higher education establishment has to invest a huge amount of money on the licensing of these applications. One of the alternatives is to avail cloud-based services for teaching multimedia applications in higher education establishments. This research paper studies the benefits of using cloud computing for multimedia-based applications.

A. Z. Bhat (✉) · V. R. Naidu · B. Singh
Vivekananda Global University, Jaipur, India
e-mail: anjum_zameer@hotmail.com; azameer@mec.edu.om

V. R. Naidu
e-mail: vikasrn@gmail.com

B. Singh
e-mail: baldev.vit@gmail.com

A. Z. Bhat · V. R. Naidu
Middle East College, Muscat, Oman

© Springer Nature Singapore Pte Ltd. 2019

V. S. Rathore et al. (eds.), *Emerging Trends in Expert Applications and Security*,
Advances in Intelligent Systems and Computing 841,
https://doi.org/10.1007/978-981-13-2285-3_81

drksujatha23@gmail.com

1 Introduction

Cloud computing has revolutionized the world of information and communications technology by providing various IT services like a utility. A few years back, IT and computing resources were never thought to be something that can be hired or taken as a utility like the electricity, water and telephone; however, we have seen from past several years now that information technology infrastructure, i.e. hardware, software, network, applications, operating system and various other components are provided as a service. Cloud computing in its inception was available with basic three service models and over the period, these service models have extended into many other service models like network as a service, virtual private network as service [1].

As the implementation and use of cloud, computing is increasing in many organizations and educational establishments certainly not being the exception. Many researchers are working towards appropriateness of the service models for a specific type of the enterprise. A lot of research has been done to introduce a service model that is appropriate for the education establishments [2]. Many of the service models emphasize the requirements for certain specific services that are suitable for a particular type of the organization. As cloud computing becomes more and more prominent and prevalent, we may witness various alternate service architectures for different types of organizations. Multimedia applications that are in use in higher education establishments significantly increase the annual budget due to high-cost license renewals, upgradation of hardware configuration to meet the latest requirements, etc. With the implementation of cloud services, it will be more feasible for educational institutions to have regular updates, since it will cost quite lesser than the actual multimedia lab establishment costs. The migration of multimedia lab services to students benefits in multiple ways, it provides the monetary benefits to the educational establishments; however in addition to that, there are various benefits which are provided to the educational establishment as well as the students who are utilizing the services. Following are the significant benefits that are provided by using cloud services for multimedia applications:

i. Upgradation of multimedia applications on a continuous basis is not the responsibility of the educational institution; however, it happens through cloud service provider.
ii. Minimal infrastructure required at the institution for student to access any type of multimedia service.
iii. High-end multimedia applications access with the use of low-end computing resources.
iv. Very nominal subscription amount for the use of multimedia resources in an easy pay-as-you-go manner.
v. Huge monetary benefits for the educational establishment.
vi. No administrative overhead for the maintenance of high-end multimedia labs.

drksujatha23@gmail.com

The use of cloud computing for multimedia applications would be more fruitful and much more compatible if slight modification is brought to the basic cloud architecture. The cloud comes in three basic architectures those are Infrastructure as a Service (IaaS), Platform as a Service (PaaS) and Software as a Service (SaaS). The below diagram shows the typical cloud architecture used for most common cloud implementations.

Common cloud architecture provides various services to the organizations requiring infrastructure, platform or applications. There can be one improvement to the cloud architecture so as to make it fully suitable for multimedia application services to the students of the higher education institutions. The Software as a Services (SaaS) can be further differentiated so that a separate service list can emerge for multimedia applications. This would benefit both educational establishments and the service providers in a manner that service providers can specifically provide the Multimedia as a Service (MaaS); it would be very convenient for educational establishments to opt for the appropriate cloud service most suitable for their needs. The Required cloud architecture would be as follows.

Figure 1 [3, 4] shows the service architecture with multimedia as a service. The multimedia-related software services can be amalgamated to yield a specific service model that would be suitable for the educational establishments. The cloud service for multimedia eradicates the need for the students to save their work or projects on the local drives and one more important feature is that they access their work or services of the multimedia cloud literally from anywhere they can connect to the Internet. This is on more significant development as students in most of the educational establishment find the resources occupied and are not able to work on their work continuously. The multimedia cloud implementation would eliminate the requirement for students to be present in the multimedia for doing any work related to their module curriculum or any project work.

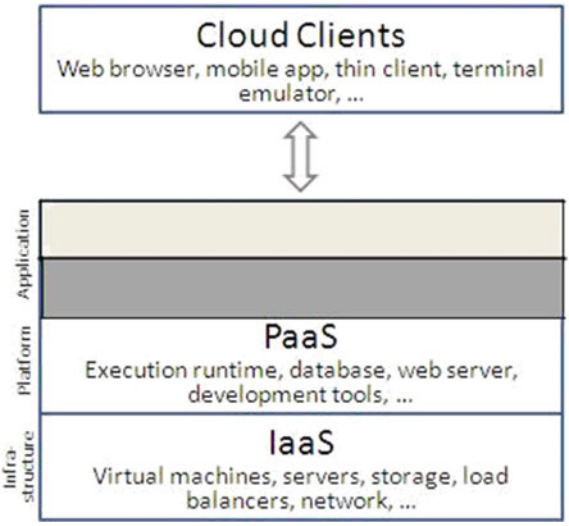


Fig. 1 Multimedia as a service

drksujatha23@gmail.com

2 Review of Current Studies

There are multiple types of cloud services, such as public cloud, private cloud and hybrid cloud [5]. Where private cloud will require the organization to provide the necessary infrastructure for the establishment of cloud services, public cloud may be chosen for such cases. In some cases, the establishment of private cloud might cost very high due to the involvement of very high-end networking devices to support seamless data streaming over the network. On the other hand, public cloud could also be a solution, where the services could be purchased online, and the multimedia-based applications could be accessed through it. One of the most famous multimedia software developers, Adobe System Incorporated, has already started their cloud services Adobe Creative Cloud (Adobe CC) and proved to be very successful, compared to their earlier standalone versions.

Earlier Adobe has released their software packages, specifically meant for educational establishments, with the name Adobe Educational Suite (ES). This suite included some basic and easy to learn tools by the educators, to develop educational content. But the last version was released in the year 2012 and later, many other software vendors have started developing multimedia-based application development tools and some of them are based on cloud as well.

Hybrid cloud is another solution for educational establishments where a part of cloud services could be taken from public cloud, and the important part which contains some confidential data of the institution could be retained on private cloud [5]. This could also be a one-time investment, as the dynamic part of the cloud, in which multimedia tools are accessed, could be taken from public cloud. The main advantage of this would be the regular updates of the tools, which will be done by the respective software vendor on the cloud itself and the education institution just need to take care of the access of those resources, at their campus, by means of good infrastructure of the network.

These days many educational institutions are moving towards enhancement in their teaching and learning strategies, and at the same time, they must look for such scenarios which are feasible to implement due to budget, when it specifically comes to multimedia software. Hybrid cloud is one of the best solutions for the institutions to implement for multimedia studies since there are many free to use multimedia resources, available to use for educational purposes. Institutionalizing some of these referred resources could also be very beneficial for the faculty as well as the institution.

The concept of Massive Online Open Course (MOOC) is based on this, and millions of users are taking advantage of such courses; some of them are free and some are available with nominal charges to be paid. On completion, some of these MOOCs are offering course completion certificates. Some successful MOOCs are Khan Academy, Coursera and Udemy.

The learning needs of present students are entirely new and different as compared to the ones couple of decades ago. A student of this generation prefers to use more technology and mobile-based application, which are handy and time-saving

drksujatha23@gmail.com

[6]. A cloud service provides a personal workspace to the end users, whether it may be faculty or the student. Apart from these, it increases interactivity in the teaching and learning process. Moreover, since most of this multimedia-based work is carried out on cloud, there is no need for regular backup or maintenances.

It has always been a good idea if more than one education body collaborates virtually with another one to contribute to a knowledge pool. This could lead to many shared resources mutually among the establishments, which further will decrease the cost of producing learning materials along with shared resources [7]. There are several free applications, which are ready to use to produce shared resources. Many of such services are offered even on the Google cloud. Not only this but also various cloud storage has made the process even easier by offering up to 5 GB of free space which is good enough for sharing educational materials such as electronic books, quizzes, surveys, etc.

On the other hand, some of the paid cloud services offer full flexibility of customization of resources, apart from increased storage space. In certain cases, the educational institution has their own learning management system, but due to budget-related constraints, they are unable to implement on private cloud. In such a case, they opt for outsourcing the cloud service to public cloud vendors.

Over the period, cloud services in education sector have proven increase in productivity of the students by online collaboration, conferencing and instant support. As per the statistics provided by IBM which studied the articles, Edudemic reports by the end of 2014 [8],

- 68 percent of institutions use (or will use) the cloud for conferencing and collaboration,
- 65 percent of institutions use (or will use) the cloud for storage,
- 65 percent of institutions use (or will use) the cloud for office and productivity suites, and
- 62 percent of institutions use (or will use) the cloud for messaging.

The private section of the hybrid cloud which could cache the important and most frequently browsed educational content could be very time saving on repeated and multiple accesses by the staff and the students at the same time. In conjunction with an in-house learning management system, in which the confidential information of assessments could be stored, the hybrid cloud implementation could be very useful. In certain establishments, there is an implementation of their own learning management system, such as Moodle, which facilitates the process of providing the learning resources to the students with an easy and customizable interface. The flexibility and dynamism of implementing cloud services for multimedia make this as the best option for the education sector for current scenario of teaching and learning process in higher education establishments.

drksujatha23@gmail.com

3 Protection for Cloud-Based Multimedia Content

When the educational establishments are moving towards upgrade in their infrastructure to house the latest technology and trends in teaching and learning, the protection of content is another big responsibility. Protection of cloud-based multimedia content, where large-scale content development is planned, could be done by a few recommended methods [9].

One of the simplest methods is watermarking. As we know that most of the multimedia-based content on cloud consists of videos and animations, adding watermark could be a good solution to ensure that the original author's name could not be taken out of the video. Also, the developer should include a disclaimer at the beginning of the video that it is intended solely for educational purpose of that organization, and usage outside the organization is to be done after obtaining prior permission from the authors. Still, it is very difficult to control the duplication of this educational content on Internet, due to the availability of lots of editing tool, which even sometimes masks the watermark itself and can add totally different credit roll.

Another recommended method is the implementation of digital signatures methods using fingerprints on 3D videos while the author is uploading them on the streaming server/cloud [9]. With these duplicate uploads could be avoided for the similar video content.

4 Reflections on Current Practices and Recommendations

Currently, most of the multimedia applications that are used in higher education establishments are based on the in-house infrastructure that is developed to support the multimedia modules curriculum and the pragmatic concepts related to the multimedia modules. Most of these labs are equipped with high-end hardware and expensive software applications which need renewal of licenses after the expiration. The hardware components are also upgraded within 3–5 years so as to match the requirements of new applications. Due to high expenses involved in establishment of these labs, most of the educational establishments are having limited number of labs with restricted or limited applications available for students to practice. Multimedia concepts need pragmatic exposure so that students can comprehend the concepts and can deliver when it comes to working in a professional environment. It becomes mandatory for educational establishments to provide the required underpinning of practical components of multimedia irrespective of the price tags these applications may carry. One of the effective ways of achieving that would be to migrate the multimedia labs to the cloud infrastructure and take multimedia-based services for students. This would provide students with comprehensive application access in a holistic and broader perspective apart from

drksujatha23@gmail.com

various other features and flexibilities. Migration of current multimedia service to the cloud infrastructure in highly recommend for higher education establishments to achieve benefits that are manifold.

5 Results

Cloud computing also referred as on-demand computing has revolutionized the IT world. Higher education establishments need to yield extraordinary benefits offered by cloud computing. Certain areas in higher education like multimedia specialization can migrate the multimedia services provided to the students to a public cloud. This would facilitate students with enhanced feature, flexibility and ease of access apart from providing them access to such multimedia applications which are very expensive. The benefits of cloud-based multimedia facility are very much the need of the time for higher education establishments having multimedia subjects in their academic curriculum. There are enormous and diverse benefits offered by cloud computing, the changes in cloud service architecture can bring appropriate service list yielding more apposite benefits to the users. This research paper provides a reflection of the benefits that would be acquired by educational establishments as well as the enormous benefits and facilities to the students and the flexibility of access from anywhere and anytime.

Acknowledgements We are first thankful to Almighty God for bestowing us the capability to complete this research work. A word of thanks is due to all those who helped us in completing this research work including friends, colleagues, associates and experts in the field apart from our family members who always encouraged us. We are thankful to Middle East College for supporting the research activities and encouraging us always for our research work. Thanks are also due to Vivekananda Global University for their continuous support and cooperation.

References

1. Bhat AZ, Khalfan D, Vikram A (2016) Virtual private network as a Service. IEEE Explore, Delhi
2. Zameer A, Singh B (2017) Learning resources as a service (LraaS). IEEE Explore, Noida
3. Duan Y, Fu G, Zhou N, Sun X, Narendra N, Hu B (2015) Everything as a service (XaaS) on the cloud: origins, current and future trends. IEEE
4. Mell P, Grance T (2011) The NIST definition of cloud computing. National Institute of Standards and Technology, U.S. Department of Commerce
5. Singh U, Baheti PK (2017) Role and service of cloud computing for higher education system. Int Res J Eng Technol (IRJET) 04(11):708–711
6. Almajalid R (2017, June 4) A survey on the adoption of cloud computing in education sector. Computers and Society, pp 1–12
7. Naik AB, Ajay AK, Kolhatkar SS (2013) Applicability of cloud computing in academia. Indian J Comput Sci Eng (IJCSE), 11–15

drksujatha23@gmail.com

8. Tate AR (2014, Auguest 8) Five ways cloud is enhancing higher education. Retrieved 15 Jan 2017, from https://www.ibm.com/blogs/cloud-computing/: https://www.ibm.com/blogs/cloud-computing/2014/08/five-ways-cloud-is-enhancing-higher-education/
9. Niharika M, Sahoo PK (2016) Protecting cloud based multimedia content using 3-D signatures. Int J Adv Comput Tech Appl (IJACTA) 4(1):205–208

drksujatha23@gmail.com

Optimal Multi-document Integration Using Iterative Elimination and Cosine Similarity



Fr. Augustine George and M. Hanumanthappa

Abstract Summarization holds vital significance in the age of a data-centric world. This research paper tries to retain an overall summary of the article using an optimal multi-document integration. Nevertheless, location-based extractions are biased, and the data loss is high compared to the proposed approach that derives a matrix subset from a random permutation of elements after which extraction takes place. The proposed extractive method of integrating multi-documents has two stages. Initially, an iterative elimination of matrix subsets is done, finally the documents are integrated, and their lexical similarity is computed. The retention rates recorded in this implementation are considerably high even when the compression rates are increased. As proof of concept, our experiment results of extractive integration reveal high retention rate that could improve the quality of the generated summaries.

1 Introduction

Automatic keyword extraction is the process of selecting words and phrases from the text document that can project the core concept of the document automatically [1]. Text summarization can be done two ways, namely, abstractive summary and extractive summary. In abstractive method, documents are summarized based on lexical meaning that uses topic modeling and concept hierarchies. Extractive summarization methods generally identify significant sections of the document and generating summary. Abstractive summarization interprets and inspects the text using natural language processing in order to generate a short text that portrays the most important information from the input text. Researchers proved extractive

Fr. A. George (✉)
Kristu Jayanti College (Autonomous), Bharathiar University, Bangalore, India
e-mail: augustine@kristujayanti.com

M. Hanumanthappa
Bangalore University, Bangalore, India
e-mail: hanu6572@hotmail.com

© Springer Nature Singapore Pte Ltd. 2019

V. S. Rathore et al. (eds.), *Emerging Trends in Expert Applications and Security*,
Advances in Intelligent Systems and Computing 841,
https://doi.org/10.1007/978-981-13-2285-3_82

drksujatha23@gmail.com

summaries often give better results compared to automatic abstractive summaries [2]. There is a rising need for effective information retrieval tools to assist in organizing, processing, and retrieving the pertinent information and portraying them in a concise way and in suitable user-friendly format. This research addresses the problem of integrating multiple documents using iterative matrix subset elimination before abstractive summarization.

Our proposed approach to extractive integration consists of the following two phases: **iterative elimination of matrix subsets and lexical similarity computation**. The rest of the paper is organized as follows. In Sect. 2, we give an overview of text summarization related work. The next two sections describe proposed matrix subset elimination. Section 5 suggests the discussions on the obtained values. Section 6 presents venues for future researches, and Sect. 6 gives conclusions and directions for future work.

2 Related Works

Culmination of text summarization was started in 1950s. Due to the lack of improvement in natural language processing (NLP), early work of text summarization decided to focus on text genres like sentence position and cue phrase [3]. Authors [4–8] have tried text summarization using artificial intelligence from 1970s to early 2000s. Many researchers focused on identifying conceptual entities and extract relationships between entities with different inference mechanisms. Due to the exponential growth of the information storage and retrieval, retrieval engine [9–14] employed extractive text summarization methods. Similar to the tasks of information retrieval, text summarization was also done by identifying significant sentences from a document. However, most of the extractive text summarization systems are limited to single documents and IR techniques that have been exploited using extractive in text focus on symbolic-level analysis and not consider semantics such as synonymy, polysemy, and term dependency (Fig. 1).

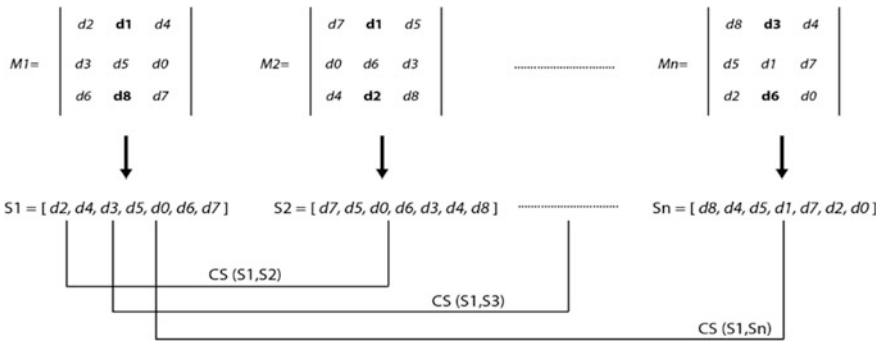


Fig. 1 Random matrix and subset generation

drksujatha23@gmail.com

Table 1 Extractive methods

Tf-Idf	Clustering	Machine learning	Fuzzy logic	Graph-theoretic approach
Bag-of-words (BoW) model. Useful in lexical level feature and does not capture position in text, semantics, co-occurrences	Requires lot of training and not suitable for small datasets. Redundancy, data loss	Reinforcement learning. Not susceptible to changes	Cue feature and location-based extraction. Lot of human intervention at each phase	Coherent chunks by a set of rules that for adjacent sentences. Lower accuracy and relevance measures

2.1 Extractive Summarization Methods

Extractive summarizers aim at picking out the most relevant sentences in the document while also maintaining a low redundancy in the summary. Some of the extractive methods discussed below in Table 1.

3 Proposed Methodology

Till recently, location-based algorithms have been formulated to pick out the most impactful statements in a document. They are often very biased as they tend to stick to extracting a particular pattern of statements. In our proposed methodology, equal priority is given to every sentence in the document. Automatic integration of multi-document using iterative matrix subset and cosine sentence similarity is an extractive method using iterative elimination depicted in Fig. 2. As we are currently focusing mainly on an extraction method rather than an abstractive one, we tokenize the sentences instead of tokenizing them into words. The application of lexical similarity through Parts of Speech (POS) tagging has applied to dataset through the algorithm and that led to a more optimized and efficient output. It focuses on computing a summary, which removes lexically similar sentences/features of the document, including the unique feature of the topic discussed. This algorithm is a promising one because (1) it is a balanced combination of static and dynamic methods and (2) no much human intervention is needed.

drksujatha23@gmail.com

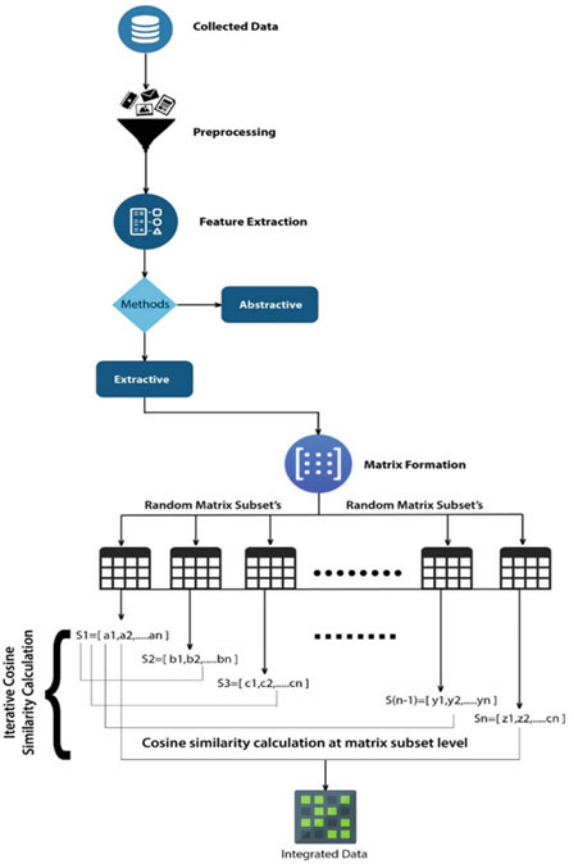


Fig. 2 Automatic integration using iterative elimination

4 System Design

Our research motive is to propose a new extractive method to integrate multiple documents and extract relevant abstract using iterative elimination of matrix subsets and cosine similarity. For testing the abovementioned method, we have used a dataset which has 16 observations and 2 attributes. These documents are preprocessed and normalized before applying the model.

The overall methodology is explained in the following steps. For explanation, we have used a simple dataset consisting of 9 documents; each document is numbered as 1–9 and is represented in a 3×3 matrix. The size of the matrix is determined based on the number of documents need to be summarized.

drksujatha23@gmail.com

4.1 Steps

Step 1: **Data collection and building the corpus**
Find the specific datasets/reviews.
Our methodology is experimented on the following dataset reviews.csv; it has 16 observations and 2 attributes.

Step 2: **Preprocessing and tokenization**
Data obtained during the data-collection phase is often inconsistent and incomplete. Data preprocessing involves extraction of vital elements from unstructured text data. In our extractive document summarization, we have done the removal of stop words and tokenization.

Step 3: **Iterative matrix subset formation**
All documents are collectively arranged randomly in a tabular format whose dimension is determined by the number of documents and these matrices are used to get the subsets.

Step 4: **Subset computation**
Arrange the documents as a matrix in the following format.
Elements are computed by mapping the length of the document to its nearest perfect squares and no_cols = SQRT (elements). The input dataset is represented as follows:

$$\mathrm{D} = \{\mathrm{d1}, \mathrm{d2}, \mathrm{d3}, \ldots \mathrm{dn}\} \Rightarrow$$

Based on the number of sentences in the document Di, the square matrix size is fixed after generating NA values in the subsequent empty rows; post that the random numbers that act as sentence IDs are generated. Iterative elimination is performed on the random matrix to obtain the subsets of each random matrix.

Step 5: **Computing cosine similarity**
We perform cosine similarity between two each subsets in a permutation. Cosine similarity defines the dot product of vector representation of individual documents (1). The documents or paragraphs represented as vectors; matrix also being one form of vector can implement this feature. The formula is as mentioned below:

$$sim(a, b)\frac{a * B}{|a| * |b|} \tag{1}$$

Calculated cosine similarity value of combinations of subsets represented in the above Table 1. The threshold value defines how the similar the subsets are to each other. If the cosine similarity of two subsets (S1 and S2) is more than the defined threshold value, then the first subset is included in the input of integration and S2 is excluded.

drksujatha23@gmail.com

Step 6: **Sorting** the sentences using their sentence IDs, thus giving a correlation and a better flow in reading the generated summary.

Step 7: **Integrate the data**
Cosine similarity is applied to the integrated input that is derived from the previous step. Once again, the same process is performed on the combinations of the different documents on the final subset of data. Finally, we get a summary with retention rates with respect to the compression rate of that document.

5 Results and Discussion

We have evaluated our approach in the dataset consists of 16 documents, and results of automatic integration multi-documents are shown in the graph. Figure 3 shows the cosine similarity values of all subsets extracted from the iterative elimination. Unique subsets are identified and integrated into the table. The ML algorithm in the above example found that subsets that possess values greater than 0.25 can be eliminated. One huge challenge when it comes to summarization is data loss. The algorithm considered optimal if it retains the vital information while eliminating the repeated or unnecessary information.

However, what is important and what is not is just perception. We can, however, note that summaries do not give details about a subject; they just reveal the onshore concepts and idea of the article. In this paper, an attempt was made to overcome bias while also sticking to the generation of a traditional summary.

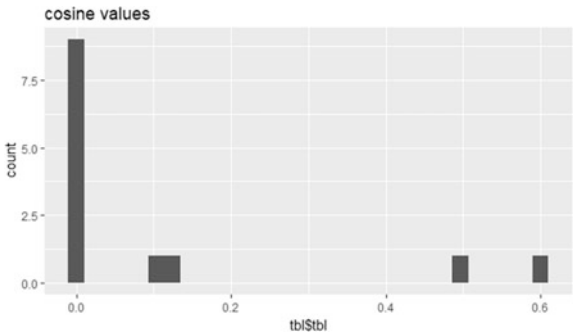


Fig. 3 Cosine values of extracted subset

drksujatha23@gmail.com

6 Conclusion Future Enhancement

The proposed method was tested against a dataset containing 16 documents and each document containing at least 10 paragraphs. In the first iteration 15% of document, paragraphs are generally eliminated and in the round of computing the lexical similarity, 30% of the document paragraphs removed. Finally, the documents are integrated and another round of the same process begins to remove redundancy. Finally, we are able to achieve at least 30% of compression and 80% retention. The integrated document was summarized using Latent Dirichlet Allocation (LDA) method in future.

References

1. Bharti SK, Babu KS, Jena SK, (2007) Automatic keyword extraction for text summarization: a survey. Eur J Adv Eng Technol (EJAET) 4(6):410–427
2. Erkan G, Radev DR (2004) LexRank: graph-based lexical centrality as salience in text summarization. J Artif Intell Res (JAIR) 22(1):457–479
3. Luhn HP (1958) The automatic creation of literature abstracts. IBM J Res Dev 2(2):159–165
4. DeJong GF (1979) Skimming stories in real time: an experiment in integrated understanding. Doctoral Dissertation. Computer Science Department, Yale University
5. Graesser AC (1981) Prose comprehension beyond the word. Springer-Verlag, New York
6. Shank R, Abelson R (1977) Scripts, plans, goals, and understanding. Lawrence Erlbaum Associates, Hillsdale, NJ
7. Young SR, Hayes PG (1985) Automatic summarization and classification of banking telexes. In: The second conference on artificial intelligence applications: the engineering of knowledge based systems. IEEE Computer Society Press, DC, pp 402–408
8. McKeown K, Radev DR (1995) Generating summaries of multiple news articles. In: Proceedings of the 18th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR'95), pp 74–82, Seattle, WA, USA
9. Kupiec J, Pedersen J, Chen F (1995) A trainable document summarizer. In: Proceedings of the 18th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR'95), pp 68–73, Seattle, WA, USA
10. Hovy E, Lin CY (1997) Automatic text summarization in SUMMARIST. In: Proceedings of the ACL'97/EACL'97 workshop on intelligent scalable text summarization, pp 18–24, Madrid, Spain
11. Mani I, Bloedorn E (1999) Summarizing similarities and differences among related documents. Inf Retr 1(1–2):35–67
12. Salton G, Singhal A, Mitra M, Buckley C (1997) Automatic text structuring and summarization. Inf Process Manag 33(2):193–207
13. Gong Y, Liu X (2001) Generic text summarization using relevance measure and latent semantic analysis. In: Proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR'01), pp 19–25, New Orleans, LA, USA
14. Yeh JY, Ke HR, Yang WP (2002) Chinese text summarization using a trainable summarizer and latent semantic analysis. Lecture notes in computer science, vol 2555. In: Proceedings of the 5th international conference on Asian digital libraries (ICADL'02), pp 76–87, Singapore. Springer-Verlag, Berlin

drksujatha23@gmail.com

Secured Data Sharing in Groups Using Attribute-Based Broadcast Encryption in Hybrid Cloud



E. Poornima, N. Kasiviswanath and C. Shoba Bindu

Abstract Data sharing in the cloud has several benefits to the organization as well as the users. With the advent of cloud services, Google Docs provides data sharing, such that the users can share the documents and collaborate with the other users effectively. In this paper, we addressed the challenging security issues associated with the data sharing in the cloud. Broadcasting is an efficient way of sharing files securely to several receivers. In our proposed work, we proposed a solution based on Attribute-Based Broadcast Encryption, secure data sharing in groups using Attribute-Based Broadcast Encryption, which is both secure and efficient. The importance of data sharing and the necessity to ensure security and privacy is discussed with existing literature. Review of existing methods of achieving data sharing in the cloud is discussed. The primary difference between the proposed solution and traditional broadcast encryption is that security is achieved dynamically based on the parameters supplied during the encryption and decryption process. Our solution allows selecting and revoking the users based on the attributes. We propose three algorithms, namely, Secured Encryption using Group-Based Signature Generation (SEGB-SG), Secure Verification using Group-Based Signature Verification (SVGB-SV), and Revocation List based on Group Signature (RL-GS). The results for the proposed solution show a significant improvement in providing security for data sharing in the cloud. The experiments were conducted and the results are compared with the existing work on KP-ABE, and the proposed solution shows that our solution is both secure and efficient.

E. Poornima (✉)
G. Pulla Reddy Engineering College, JNTUA, Kurnool, India
e-mail: poornimacse561@gmail.com

N. Kasiviswanath
Department of CSE, G. Pulla Reddy Engineering College, kurnool, India
e-mail: hodcse@gprec.ac.in

C. Shoba Bindu
Department of CSE, JNTUA, Anantapur, India
e-mail: shobabindhu@gmail.com

© Springer Nature Singapore Pte Ltd. 2019

V. S. Rathore et al. (eds.), *Emerging Trends in Expert Applications and Security*,
Advances in Intelligent Systems and Computing 841,
https://doi.org/10.1007/978-981-13-2285-3_83

drksujatha23@gmail.com

1 Introduction

The cloud computing is defined as "a model for enabling ubiquitous access to describe trusted third party hosted services". The advantages of cloud services include scalability, accessibility, collaborative, reliability, and low cost. Agriculture in Indian economy plays a major role. According to the survey, more than 70 percent of the country of rural households relies on agriculture. With the advent of latest technologies in the agriculture filed such as smart farming, Internet of Things, and cloud computing, there is a strong need for secure data sharing [1, 2]. Figure 1 shows the basic cloud model.

Cloud computing [3] in agriculture provides a means for accessibility and affordability, which helps the farmers to share the information related to crops and can help in precision farming [4].

Figure 2 describes the various types of clouds such as private, public, and hybrid. Cloud services can be used in the agriculture field to monitor the crop fields and capture values such as temperature, CO_2 concentration, moisture, and precipitation. Data sharing in cloud has become an important service in cloud storage. The primary goal of this work is to securely and efficiently share data with other users in the cloud storage in a flexible manner. Cloud services can be used to provide secure data sharing capabilities. The agriculture field has been driven by technological development and innovation.

Amazon-like cloud service providers enable to deliver a variety of services to their users through the central data centers. By sharing the local data management systems in the servers, clients are provided with rich quality of services and can protect consequential investments for the native infrastructures. However, it also has several risks and threats to the privacy and confidentiality of the stored information in cloud. This paper presents secure data sharing related to agriculture data stored in the cloud, for dynamic groups based on attributes such as geographic location, area, farmer id, etc. This paper also discusses encryption based on the attributes, which can also be used for broadcasting the data if needed to share the vital information related to crops.

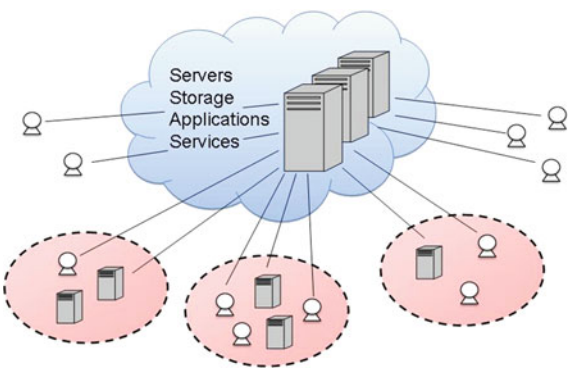


Fig. 1 Basic cloud computing system

drksujatha23@gmail.com

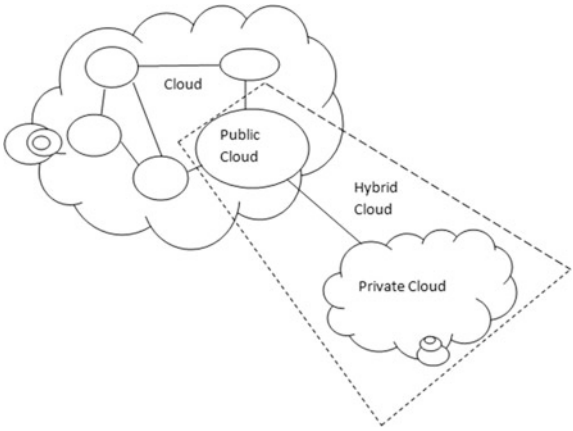


Fig. 2 Types of cloud

A broadcast encryption scheme [5] is useful in several public and commercial applications such as secure data sharing and broadcast of multimedia. Dynamic broadcast encryption [6] enables encrypted data to be transmitted to the group of users such that only authenticated users are allowed to decrypt the data. Attribute-based encryption has been introduced in [7, 8], which encrypts and shares the encrypted data which fulfills certain prerequisites.

The Identity-Based Encryption (IBE) is a technique in which a public key is considered to be an identity (ID) of a user. The definitions indicated below represent the basic preliminaries for the proposed solution.

- Consider two groups say G_1 and G_2 of size q, where q refers to a prime number. IBE uses bilinear mappings $\tilde{e}\, G_1 = \times\, G_1 \rightarrow G_2$. The mapping must satisfy the following properties of non-degeneracy and computable.

A mapping is defined as $\tilde{e}$: $G_1 \times G_1 \rightarrow G_2$ is bilinear if and only if there exist $\tilde{e}(xP, yQ) = \tilde{e}(P;Q)^{xy}$ for all $P, Q \in G_1, \forall x, y \in Z$, *is called as Bilinear map. A bilinear mapping which satisfies the above three properties is said to be an "admissible bilinear map".*

To protect the confidentiality of agricultural data in cloud, secure encryption based on attributes is used, which needs exchange of decrypted keys between the group users and the owner of the data. Because of exchanging the keys through the service provider [9, 10], there is risk of violating the privacy.

As the clouds are being more sophisticated, there is a need for secured data sharing to be provided in the cloud services related to sharing data files across several users with their respective groups [11].

drksujatha23@gmail.com

2 Related Work

Cloud computing has several advantages and also poses numerous challenging security issues [12]. The use of the cloud for agriculture data, including crop related parameter, rainfall, temperature, and crop yield, should be stored and shared securely. The current cryptographic primitives for secured data sharing in cloud computing are not directly used because the owners will lose control of the data.

Hence, there is a need to provide a perfect data storage security mechanism. Second, cloud computing cannot be completely relied on trusted third party. To ensure data sharing security under dynamic group environment needs an efficient and a novel secure mechanism.

Kallahalla et al. [13] made a proposal of a solution to cryptographic storage which provides sharing of files in a secure way without trusted third party on the file servers in the cloud. The no. of cryptographic keys distributed between the users is reduced for file read, write, groups and user revocation, which has high overhead as compared with the existing literature.

KP-ABE technique was proposed in [14] for cloud storage security, where each user is given two keys: the first one is named as group signature key and the other one as attribute key. Thus, attribute-based encryption is made possible.

An architectural framework of rule-based expert system was designed and developed by Kaliuday Balleda, which can manage rice and wheat crop pest. The symptoms caused by insects in rice crop and wheat can be diagnosed by this expert system. This expert system has user interface interactive console for analyzing of responses made by the client against the queries related to the particular symptoms of the diseases. This framework is useful in predicting diseases based on the information and symptoms given by user.

Vimercati et al. [15] proposed key derivation methods [16] for securing data storage over untrusted servers, a symmetric key is used for encryption of each file, and the secret key is provided for user. To provide authorizations for the privileged user, the data owner generates respective tokens in public along with corresponding secret key. The user can obtain the data using decryption keys of desired files.

According to the authors [17], the following issues are to be addressed related to data sharing in the cloud servers:

Access to stored data in cloud—Unauthorized users, wicked process, and malicious virtual machines must not be allowed to access the data in the cloud. Nowadays, cloud service providers are using a few type of encryption techniques to guarantee access for only authorized users. But there exists a real issue with the current frameworks that there is an exchange off between extremely secure information and computational overhead. As services are offered on "a pay-as-you-go" bases, we need to prevent encryption and decryption operations to hold cost down. Our existing traditional encryption systems today are not effective to use for cloud computing technologies, particularly while we consider large volume of data is to be stored in the cloud.

drksujatha23@gmail.com

Data segregation and storage location—In businesses, users do not want to store data at the server side. A few countries have in place laws relative to the location of user data. Moreover, only a few cloud service providers acquire local record storage laws (it depends on the client to ensure compliance).

Data availability and backup—Generally, user wants to access their data from any location, any time, even though service providers may be offline. They additionally need their information to be backed up regularly. Not all providers currently mirror their data storage widely enough.

Confidentiality—The service provider cannot read or learn any information about the data.

Integrity and non-repudiation—If any modifications done by an unauthorized user must be identified by the client and access to client data is logged.

Availability—Data stored in the cloud must be available to access at any location, from any device and at any time.

3 Proposed Work

We present the architecture shown in Fig. 3 of our secure cloud storage system in the hybrid cloud. In the hybrid cloud, the private cloud stores the sensitive information associated with the group manager and users membership certificates, and the public cloud is used for data storage, to store the encrypted data, and the public parameters associated with the ABBE system. The user who wants to access the information communicates public cloud. The private cloud maintains the group manager secret key generation parameters and the key used for generating the membership certificates of the users. The authority or PKI issues the public and private keys for users and the group manager.

At the first step of the proposed solution, the setup phase is included. In this phase, the inputs are the security value, and the no. of attributes of the users in a group is taken as inputs. The algorithm produces the output as the public key

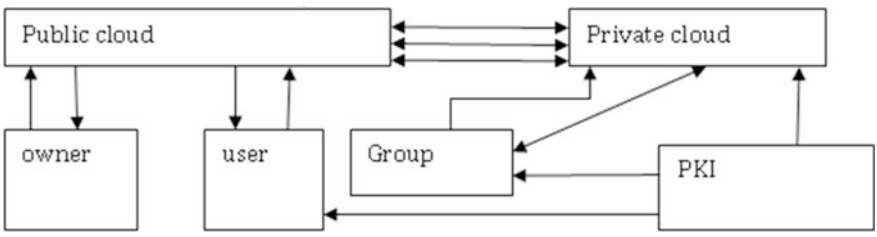


Fig. 3 Data sharing architecture in hybrid cloud

drksujatha23@gmail.com

(PK) and master key (MK), where the public key is used at the time of encryption and MK is used for generating the private key and is kept secret.

PK=(g_1, g_2, g3,…g_k, g_{k+1},…g_{2k}, v) € G_0^{2k+1} and MK={ϑ, α}

Key generation:
In the key generation, it takes PK, MK and A (users attribute list), Au = $\{A_u[i]_{i€[1,k]}\}$ as inputs and returns the users private keys.

Private key of each user, $SK_u = (D,\{D_i\}_{i€[1,k]}, \{F_i\}_{i€[1,k]})$

Encryption:
The encryption algorithm takes the input as PK, original message M and AL specified access policy with k attributes, where each k = +(positive), −(negative), * (wildcards), and returns the output as ciphertext (CT).

Ciphertext, C = E(A{,$\{M\}_{key}$, CHDR) where CHDR = ciphertext header = $\{C_0, C_1\}$

Decryption:
The decryption algorithm takes the input as PK, SK of a user and the ciphertext (CT) and outputs the original Message(M). M = D $(C)_{key}$

Group manager
Group manager is responsible for delivering many issues like registration of new users, revocation of users, and initializing system parameters; in case of any dispute, he reveals the original identity. The group manager is assumed to be fully trusted as he is acted by the third party in the given ADaas.

Group members

Group members (U_i) are the set of registered users who store private data or sensitive data in cloud and allow them to share with the others group members, and their membership is dynamically changed.

Attribute Authority

It is a trusted attribute center and is in charge of issuing the attributes to usersand generating user's attribute secret key.

- Confidentiality,
- Access control,
- Efficiency,
- Anonymity,
- Traceability,
- Non-frameability, and
- Privacy.

drksujatha23@gmail.com

User Registration

A user gets registered into the group by sending the request to the group manager using the join protocol [18] with the following parameters: (upk_i, usk_i), Ai, t. The group manager computes the private key (sk_i) and updates the registration table (reg), which contains the tuples users ui, user public key (upk_i), Secret k.

SKi, Signature (σ). The secret key consists of membership certificates and attribute certificates.

A user U_i computes the following:

- Ui picks randomly $y_1' € \in_R \mathbb{Z}_p^*$ and sends $.Y_i' = g_1^{y_1'}$. which will be known to him only and its an extractable commitment of y_i'.
- A user Ui checks the certificate issued by the group manager by computing

$$e\left(A_i, \omega\, g_2^{xi}\right) = e(h, g_2)e(g_1, g_2)^{y' + y''}$$

- If the certificate is valid, then the user computes and does the signature σ_i.

$$y_i = y_{i'} + y_{i''} \quad \text{and} \quad \sigma_i. = \text{Dsig}_{\text{uski}}\left(A_i, g_2^{xi}, Y_i\right)$$

The group manager computes the following:

- Selects $x_i € \mathbb{Z}_p^*$ and $y'' € _R \mathbb{Z}_p^*$ and computes A_i, X_i, W

  A_i= (h Y'_i $Y''_i)^{1/(a+x)}_i$, $X_i = X_{i,2} = g_2^{xi}$, $W = T_{i,j} = h^{sj/ϑ_Y + xi}$

- The private key, $Sk_i = ((A_i, X_i, y_i), \{W\}_{attj\ €Ai})$ is generated.

Here, (A_i, X_i, y_i) = Membership certificate and $\{W\}_{attj\ €Ai\ =}$ Attribute certificate.

The following steps describe the proposed work: Secured Encryption using Group-Based Signature Generation (SEGB-SG).

Step 1: It includes agriculture-specific parameters, number of users, public key, and private key, referred as public elements, which are available globally.
Step 2: The user *A* chooses a private key $R_A < Q$ and generates a public key U_A, $U_A = \alpha\ R_A$, and similarly, user *B* generates a public key U_B and private key R_B.
Step 3: *A* encrypts the text in files using M < p, which is intended for user *B* as below:

(a) Selects a random value k, where $1 \leq k \leq p$.
(b) Calculates $K_1 = (U_B)^k \bmod q$ and (C_1, C_2) where $C_1 = \alpha^k \bmod p$; $C_2 = K_1 M \bmod p$ (C_1, C_2) = signature.

Step 4: *B* checks the generated signature and verifies the signature as follows:

(a) Computes $K_1 = (C_1)^{X_B} \bmod p$, $K_1 = \left(\alpha^k\right)^{X_B} \bmod p$ $p = \left(\alpha^{X_B}\right)^k \bmod p$ $p = (Y_B)^k \bmod p$.
(b) Computes $M = (G_2 K_1^{-1}) \bmod p$, where K_1^{-1} is the multiplicative inverse of K_1.

drksujatha23@gmail.com

Therefore, $(G_2K_1^{-1})$ mod p = $(K_1MK_1^{-1})$ mod p = $MK_1K_1^{-1}$ mod p = M mod p, where G1 and G2 indicate Group 1 and Group 2.

Signature verification is the task of determining whether or not the signature is that of a given user. The following algorithm, Secure Verification using Group-Based Signature Verification (SVGB-SV), is described as follows:
Input: *K*, which is a private key and has the following form:

- a pair (*n*, *d*), where n indicates an integer ranging from 0 to 1, and d indicates the private key generated from the previous algorithm.

  Step 1: If a file *m* is not between 0 and $n - 1$, then display the output out of range and terminate the algorithm.
  Step 2: If the values n and d are provided, then follow the below sub-steps:

$$\text{Let } s = m^d \bmod n.$$

  If public key is provided as an input, then in order to verify whether the signature is authenticated or not, the following is used

$$\text{Let } s_1 = G_1(m^{dP} \bmod p) \text{ and } s_2 = G_2(m^{dQ} \bmod q)$$
$$s = s_1 + s_2. \quad \text{Output } s.$$

  The output *s*, which represents that the signature, is verified for the Groups G_1 and G_2, which is represented with the values as a range between 0 and n–1.
  The following steps describe Revocation List based on Group Signature (RL-GS):

  Step 1: Obtain the user revocation list if present from the cloud server. A user intitally adopts its secret key Sk_i, to compute a signature S on the message with groupid and fileid at time "t" by using Algorithm SEGB-SG. If the signature is verified correctly, then the local shared file is shared among the other users in the group. If the signature is not verified, then it is added into revocation list. A group user sends a data request which is having the details about (groupid, fileid, SK_i) to the cloud server. For verifying validness of a signature, a server uses algorithm SVGB-SV. After that with the help of algorithm RL-GS revocation, verification is done based upon the revocation list if it is felt necessary. As soon as the completion of a successful check, there is a response to the related data file by the cloud server and to the user the revocation list.
  Step 2: The validity of the revocation list is been verified. This procedure is identical to the step 2 of SEGB-SG algorithm.

drksujatha23@gmail.com

4 Performance Evaluation

The proposed solution ensures the key features of providing security in cloud computing such as security, authentication, data confidentiality, and authorization using secure group signature creating and verification process. In this section, we are discussing the results of the proposed work by comparing it with the existing KP-ABE solution. As the number of groups in KP-ABE increases, the time taken in secure data sharing also increases. In the proposed work, due to minimal overhead and storage, it is been made possible to securely share the file among several users in various groups.

Figure 4 describes the comparison of the proposed solution (SEGSB-SG) algorithm with the KP-ABE algorithm.

Figure 5 describes the comparison of the proposed solution with the KP-ABE algorithm. The performance of our proposed solution is more efficient and has better execution times and functionality.

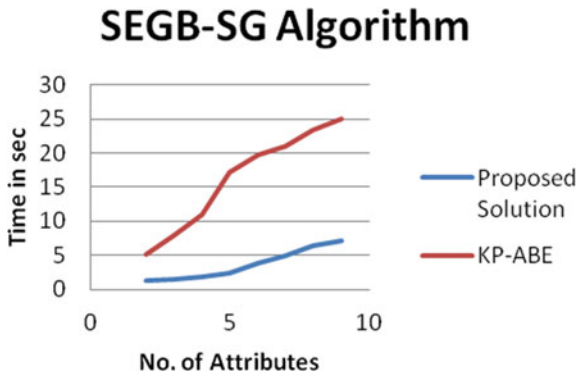


Fig. 4 Comparison of SEGB-SG algorithm

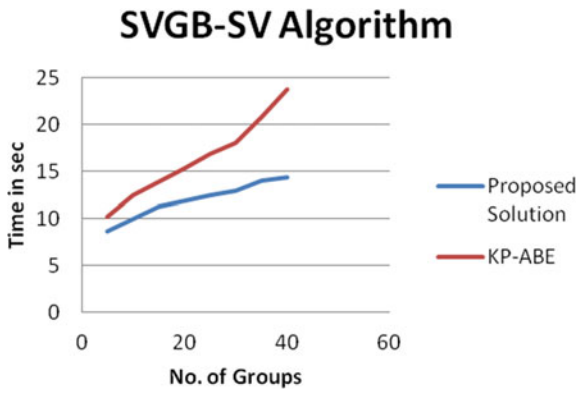


Fig. 5 Comparison of SVGB-SV algorithm

drksujatha23@gmail.com

Assessment of the performance is made using the yardsticks of communication overhead (count and size of messages), storage overhead (data stored on the users and data centers), and computation overhead (number of operations needed in encryption and decryption). During the course of communication, the system is denoted by "N" the group size. Our solution presents fast attribute revocation method that achieves both forward and backward security.

Storage Overhead

We compare the proposed solution with the original KP-ABE. The proposed solution has 6logN + 1 elements on G_0 in the public key. The storage overhead, therefore, is O(logN + m). Each ciphertext (CT) in it has exactly two group elements on G1. In G1, the size of an element is approximately 128 bytes. This is significantly smaller. Whereas in original KP-ABE technique, the user storage overhead depends on number of attributes N_a a user in the group possesses and the size of G_1. In our proposed technique, length of the group signature is constant O (1), and length of the group signature is 320 bytes, which is smaller than short group signatures. In our technique, the cost of the signature generation and signature is constant and does not depend on the number of attributes of the user, exponential pairings.

Our proposed solution is based on selective-ID attackers, collusion attacks, and Chosen Plaintext Attack (CPA).

We show our proposed scheme security analysis in terms of confidentiality, traceability, anonymity, non-frameability, and non-collusion of attributes.

Computation Overhead

In the proposed work, each encryption needs logN operations on the G_1, while 2logN + 1 pairings are needed for decryption. Here, the complexities of encryption, decryption are confined by O(logN). This clarifies that the proposed solution is more efficient, when the group size is very large and we can still notice that the cost of the computation of our technique is independent of the no. of users revoked in a group. However, in the KP-ABE scheme requires more number of modulo operations, exponential and pairing operations, as a result grows linearly with increase in no. of users in a group and the size of an access policy for the data decryption.

Moreover, our proposed technique has user privacy by concealing the access policy for the ciphertext.

5 Conclusion

Data sharing is considered as important functionality in the cloud storage. In this paper, we provide secure, efficient, and flexible data sharing with the other users in the groups. We describe a new technique, which is capable of sharing the data, with the attributes based on agriculture in groups using this Attribute-Based Broadcast Encryption technique. This technique satisfies the constraint of security and authenticity. The importance of data sharing and the necessity to ensure security

drksujatha23@gmail.com

and privacy is discussed with existing literature. Review of existing methods of achieving data sharing in the cloud is discussed. The primary difference between the proposed solution and traditional broadcast encryption is that security is achieved dynamically based on the parameters supplied during the encryption and decryption process. And we proposed three algorithms, namely, Secured Encryption using Group-Based Signature Generation, Secure Verification using Group-Based Signature Verification, and Revocation List based on attribute Group Signature in a public key cryptosystems. Besides this, the size of the ciphertext remains constant and such that an efficient d legation of decryption rights to the given set of ciphertexts is possible. In our work, a single key is compacted by aggregating a set of secret. A formal security analysis is given in standard model.

Data sharing is a key issue in big data. We extend our data sharing scheme for dynamic groups in big data for HDFS as our future work.

References

1. Sundmaeker H, Verdouw C, Wolfert S, Pérez Freire L (2016) Internet of food and farm 2020. In: Vermesan O, Friess P (eds) Digitising the industry—internet of things connecting physical, digital and virtual worlds. River Publishers, Gistrup/Delft, pp 129–151
2. Verhoosel J, van Bekkum M, Verwaart T (2016) HortiCube: a platform for transparent, trusted data sharing in the food supply chain. In: Proceedings in food system dynamics, pp 384–388
3. Khan H, Ahmad A, Johansson C, Al Nuem MA (2011) Requirements understanding in global software engineering: industrial surveys. In: International conference on computer and software modelling, Singapore
4. Chavali LN (2014) Cloud computing in agriculture. In: Kishor PB, Bandopadhyay R, Suravajhala P (eds) Agricultural bioinformatics. Springer, New Delhi. https://doi.org/10.1007/978-81-322-1880-7_12
5. Gentry C, Ramzan Z, Woodruff DP (2006) Explicit exclusive set systems with applications to broadcast encryption. In FOCUS, pp 27–38. IEEE Computer Security
6. Waters B, (2008) Ciphertext-policy attribute-based encryption: an expressive, efficient and provably secure realization. In: Proceedings of the international conference of practice and theory in public key cryptography
7. Delerablee C (2007) Identity based broadcast encryption with constant size ciphertexts and private keys. In: Proceedings of ASIACRYPT, pp 200–215
8. Sahai A, Waters B (2005) Fuzzy identity based encryption. In: Advances in cryptology EUROCRYPT 2005. Springer, pp 457– 473
9. Fiat A, Naor M (1993) Broadcast encryption. In: Stinson DR (ed) CRYPTO, Lecture notes in computer science, vol 773. Springer, pp 480–491
10. Balleda K, Satyanvesh D, Sampath NVSSP, Varma KTN, Baruah PK (2014) Agpest: an efficient rule-based expert system to prevent pest diseases of rice & wheat crops. In: IEEE 8th proceedings international conference on intelligent systems and control (ISCO), 978-1-4799-3837, IEEE
11. Poornima E, Kasiviswanth N, Bindu CS (2015) Secure data sharing for multiple dynamic groups in Cloud. In: 2015 conference on power, control, communication and computational technologies for sustainable growth (PCCCTSG), Kurnool

drksujatha23@gmail.com

12. Wang Q, Wang C, Li J, Ren K, Lou W (2009) Enabling public verifiability and data dynamics for storage security in cloud computing. In: Proceedings of ESORICS'09
13. Kallahalla M, Riedel E, Swaminathan R, Wang Q, Fu K (2003) Plutus: scalable secure file sharing on untrusted storage. In: Proceedings of the USENIX conference on file and storage technologies, pp 29–42
14. Yu S, Wang C, Ren K, Lou W (2010) Achieving secure, scalable, and fine-grained data access control in cloud computing. In: Proceedings of the IEEE INFOCOM, pp 534–542
15. di Vimercati SDC, Foresti S, Jajodia S, Paraboschi S, Samarat P (2007) Over-encryption: management of access control evolution on outsourced data. In: Proceedings of VLDB'07
16. Ateniese G, Fu K, Green M, Hohenberger S (2005) Improved proxy re-encryption schemes with applications to secure distributed storage. In: Proceedings of NDSS'05
17. Wei J, Liu W, Hu X (2016) Secure data sharing in cloud computing using revocable storage identity-based encryption. IEEE Trans Cloud Comput (99). https://doi.org/10.1109/tcc.2016.2545668
18. Blazy O, Fuchsbauer G, Pointcheval D, Vergnaud D (2011, March) Signatures on randomizable ciphertexts. In: International workshop on public key cryptography. Springer, Berlin, Heidelberg, pp 403–422

drksujatha23@gmail.com

An Efficient FPGA-Based Shunt Active Filter for Power Quality Enhancement



P. C. Naveena Shri and K. Baskaran

Abstract Maintaining power quality is being an important task within the operation of the available facility. It is compelling by the quality standards (IEEE-519) to limit the harmonics distortion at suitable intervals. The excessive use of power electronic devices in distribution system has evolved the matter of power quality; it is leading to harmonics generation and in substantial economic losses. Filters approaches are effective and economical technique for harmonics mitigation. This work proposes a novel technique with Field Programmable Gate Array (FPGA) controller for controlling the shunt active filter to mitigate the harmonics in power systems. Harmonics identification methodology and compensation management adopted are incorporated in this work.

Keywords FPGA · Harmonics · Shunt active filter · Instantaneous P-Q theory

1 Introduction

Nowadays, the energy demand increases, and the power quality is one of the key constraints in transmission and distribution. The advancements in the semiconductor technology and the proliferation of the electronic devices or the non-linear loads in distributions system cause the power quality issues.

To cope with these problems, basically, there are two approaches for the mitigation of power quality issues. Traditionally, passive filters are used to mitigate harmonics. Though it has varied benefits, it also suffer from several disadvantages such as standardization problem, fastened compensation characteristics and the size is large. The implementation of SAF has emerged as solution for reactive power compensation and harmonic damping for power distribution lines. It provides

P. C. Naveena Shri (✉) · K. Baskaran
Government College of Technology, Coimbatore, Tamil Nadu, India
e-mail: naveenacitra@gmail.com

K. Baskaran
e-mail: drbasakaran@gct.ac.in

© Springer Nature Singapore Pte Ltd. 2019

V. S. Rathore et al. (eds.), *Emerging Trends in Expert Applications and Security*,
Advances in Intelligent Systems and Computing 841,
https://doi.org/10.1007/978-981-13-2285-3_84

drksujatha23@gmail.com

effective and sensible harmonic compensation for non-linear loads. This work mainly deals with the field programmable gate array digital version of SAF, which is the necessary part of industrial control systems.

In recent times, many research works have been implemented with the digital implementation of shunt active filter for the reactive power demand and elimination of current harmonics using microcontroller unit and Digital Signal Processing (DSPs). However, DSPs or microcontroller face the problems such as large execution time, flexibility, lesser amount of accuracy, high cost and inadequate sampling period. The current controller is an essential component of SAF, in which inverter generates the controlled reference current. The current controller requires comparatively faster sampling rate, usually in microseconds, to increase the accuracy in sensing the voltage and current, which is difficult to implement in microcontroller or DSPs.

The FPGA-based real-time simulator (OPAL-RT) is used to validate the instantaneous P-Q theory, Hysteresis Current Controller (HCC) and Low-Pass Filters (LPFs). Designed control algorithm is dumped on a single all-on-chip FPGA. The proposed system has been implemented in the real-time simulator (OPAL-RT [OP4510]) under the steady-state and dynamic conditions. The outcome shows the viability of the proposed scheme in executing the process at relatively faster execution time.

2 System Design and Operating Principle

The proposed system consists of a 3ϕ SAF, connected to the grid with a non-linear load. Three-phase Voltage Source Inverter (VSI) with a DC link capacitance acts as SAF, which reduces the harmonic currents in the grid. The AC side of the VSI is connected with the series inductance at PCC. Non-linear load consists of a 3ϕ uncontrolled diode rectifier, feeding RL load. By controlling the switching of IGBTs in VSI, the compensation current has been generated and supplied to the grid. The gating signals to the IGBTs were generated on the basis of reference current, hysteresis current controllers.

3 Control Method

The shunt active filter system is based on the P-Q theory transformation, which has simple steps and circuits. When the power system was distorted, the power quality was improved by eliminating the harmonics present in the grid using this method. In addition, the system is aimed to come across the reactive power requirements of the load and to sustain the unity power factor in a 3ϕ system with unusual load conditions. The HCC generates the signals to switch the VSI, according to the reference compensation current and some other factors.

drksujatha23@gmail.com

3.1 Instantaneous Reactive Power Theory

'The Generalized Theory of the Instantaneous Reactive Power' in three-phase circuits also known as instantaneous power theory or P-Q theory was proposed by Akagi et al. in 1983. To deal with the instantaneous supply voltage and the load current, the three-phase coordinates are converted into two-phase coordinates, which is known as Clarke transformation. And the power is calculated using the instantaneous power calculation block. After that, the compensation current is calculated which is in two-phase coordinate. This can be converted into three-phase coordinate known as inverse Clarke transformation. The compensation current calculation block is given in Fig. 1.

3.2 DC Link Control

The capacitor acts as DC link, fed by the AC system. Injecting active power from the DC capacitor into the system reduces the total DC voltage. When the supply voltage is purely sinusoidal, reactive power demand is only by the current harmonics. The SAF losses are primarily due to the switching and conduction losses of IGBTs and diodes. The PI regulator is operated to generate a fundamental current

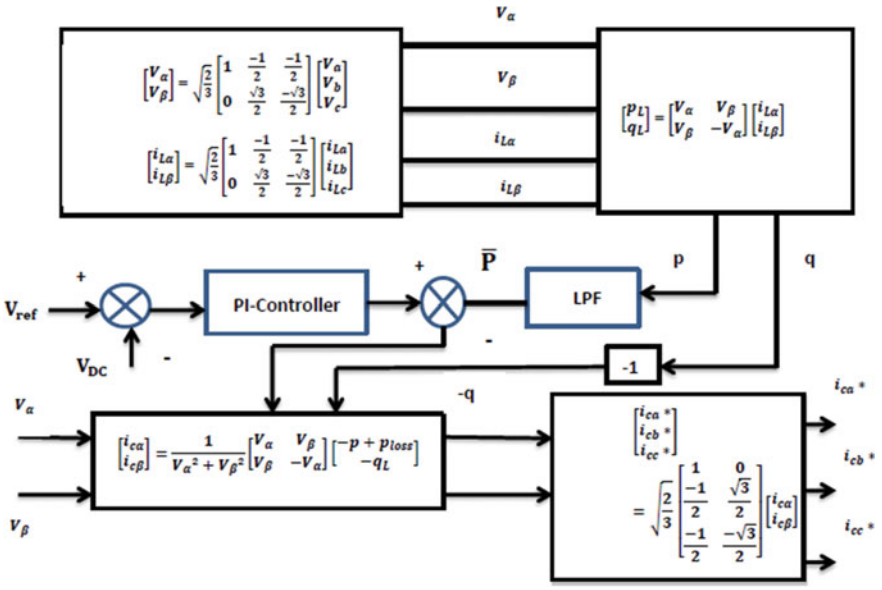


Fig. 1 Block diagram of P-Q theory

drksujatha23@gmail.com

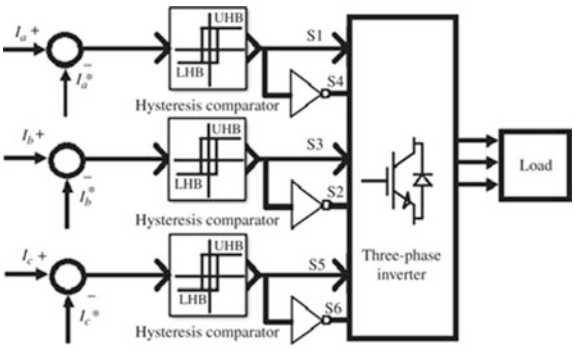


Fig. 2 Hysteresis-band current control

reference, for the DC bus voltage regulation. To control and maintain the DC link voltage, a digital PI controller has been implemented.

3.3 Hysteresis Current Controller

For the quick controllability and easy implementation of SAF, the Hysteresis-Band Current Controller (HCC) is used. In HCC, a reference value is kept and is compared with the input of the controller. Based on the error between the two inputs and the reference values, signals are generated. The VSC is controlled by the signals generated by the HCC. The hysteresis-band PWM current controller was controlled by comparing the reference active filter current $i^*(i_a^*, i_b^*, i_c^*)$ and sensed source current $i(i_a, i_b, i_c)$ to obtain the switching patterns of SAF (Fig. 2).

Here,

i_a^*, i_b^*, i_c^* = Reference compensation currents.
i_a, i_b, i_c = Actual line current of APF.

4 Simulation and Results

The model of active filter was realized through Matlab/SIMULINK. For real-time simulation, the model build in Matlab/SIMULINK must meet certain requirements: the model component blocks must be grouped in subsystems, one or more for computation (named SM_*name*) and one subsystem for user interface to display the results (named SC_*name*), and then, connection between these subsystems is realized through Opcomm block.

Master subsystem consists of calculations such as Clarke transformation, inverse Clarke transformation, PI controller and hysteresis controller, source, load and the

drksujatha23@gmail.com

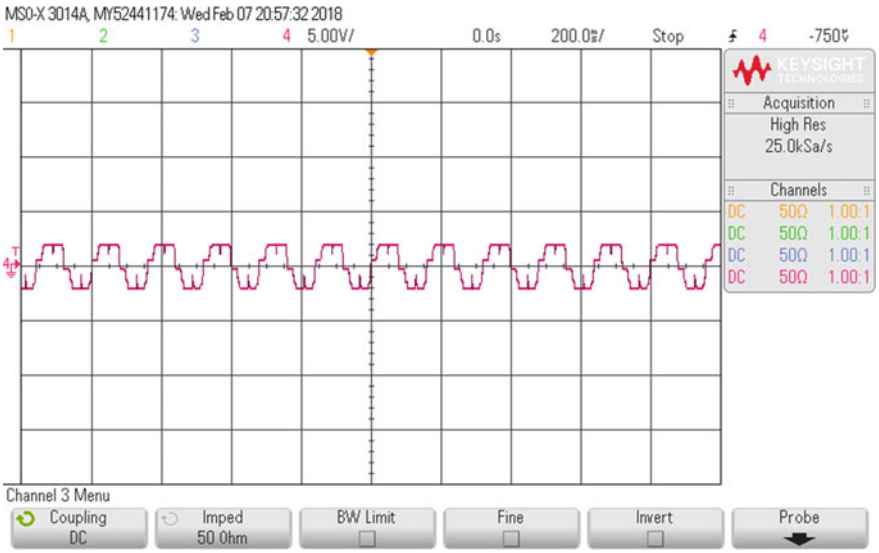


Fig. 3 Source current with compensation

inverter with the DC link voltage. Console subsystem consists of the output which is to be displayed.

The simulation was chosen for 1 µs time step, small enough to obtain results with high accuracy.

The connection between the processors is achieved through the high-speed FPGA-type board. Then, the model was loaded and compiled through OPAL-RT platform to run in real time. Hardware synchronization type was chosen to get the real-time results.

Figure 3 shows the source current with compensation. Initially, the harmonics disturbs the system, then the filter is connected to the grid and the compensation

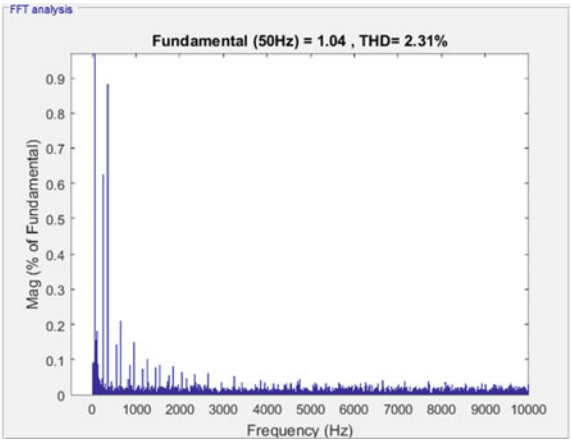


Fig. 4 THD after compensation

drksujatha23@gmail.com

current generated was injected into the grid to cancel out the harmonics. The harmonic analysis highlights the THD of load current before compensation was about 28.38% and the THD was reduced to 2.31% after connecting the SAF to the grid. This analysis shows the good filtering quality of SAF in real time (Fig. 4).

5 Conclusion

In this paper, the shunt active filter was designed with non-linear load which generates the distortion in load current. The SAF is designed using Matlab/SIMULINK to reduce the harmonics present in the load current. This simulation implementation represents the operation and control of SAF by calculating compensation current. The instantaneous P-Q theory has been used for shunt active filter, where harmonics in current have been compensated. For this proposed SAF model using real-time simulation (OPAL-RT), the current THD is reduced from 28.38 to 2.31%, which confirms the good filtering quality of current harmonics and compensation of reactive power which improves the power quality.

References

1. Shukla S, Mishra M, Singh B, Kumar S (May/June 2017) Implementation of empirical mode decomposition based algorithm for shunt active filter. IEEE Trans Ind Appl 53(3):2392–2400
2. Eskandarianet N et al (2014) Improvement of dynamic behavior of shunt active power filter using fuzzy instantaneous power theory. J Power Electron 14(6)
3. Wang Yu, Xie Y-X (2014) Adaptive DC-link voltage control for shunt active power filter. J Power Electron 14(4):764–777
4. Venkatesh Kumar P, Charles S (2012) Modelling, simulation and implementation of optimization algorithm based shunt active filter for harmonics mitigation of nonlinear loads. Aus J Elec Electron Eng 9(1):77–87
5. Singh SP, Patidar RD (2010) Digital signal processor based shunt active filter controller for customer generated harmonics and reactive power compensation. Elect Power Compon Syst 38 (8):937–959
6. Tang Y, Wang P, Choo FH, Loh PC, Blaabjerg F (2012) Generalized of high performance Shunt Active Power filter with output LCL filter. IEEE Trans Ind Infor 59(3):1443–1452
7. Liu J, Degano M, Zanchetta P, Lavopa E (2012) Control design and implementation for high performance shunt active filters in aircraft power grids. IEEE Trans Power Deliv 59(9):3604–3613
8. Monmasson E, Bahri I, Tisan A, Idkhajine L, Cirstea MN, Naouar MW (2011) FPGAs in industrial control applications. IEEE Trans Ind Inf 7(2):224–243
9. Atalık T, Deniz M, Koc E, Gercek CO, Gultekin B, Cadırc I, Ermis M (2012) Multi-DSP and-FPGA based fully-digital control system for cascaded multilevel converters used in FACTS applications. IEEE Trans Ind Inf 8:511–527

drksujatha23@gmail.com

An Empirical Study on Potential and Risks of Twitter Data for Predicting Election Outcomes



Abdul Manan Koli, Muqeem Ahmed and Jatinder Manhas

Abstract Twitter is an online news and person-to-person communication website, where users can post their thoughts and emotions as tweets concerning an issue, actuality, choice, and so on. Since its launch in 2006, Twitter has now become one of the most reputable and much publicized online tools for people to express their social and political thoughts. Since twitter enables its registered users to express their notions as tweets, which edify researcher to create methods that may utilize to foresee the result of the election based on people tweets. This paper discusses previous research works carried out in this field by different scholars related to election outcomes and tried to find the most suitable and appropriate methods in election predictions.

Keywords Twitter · Sentimental analysis · Tweets · Election · Social media Prediction

1 Introduction

Twitter as a social networking communication site propelled in 2006. It enables its users to post content and media as tweets. These tweets or short messages have initial limitation up to 140 characters, yet on November 7, 2017, the utmost multiplied to 280 characters for all dialects with the exception of Japanese, Korean, and Chinese [1]. Enrolled users can post tweets retweets and likes their favouring tweets

A. M. Koli (✉) · M. Ahmed
Department of CS & IT, Maulana Azad National Urdu University (a Central University), Hyderabad, India
e-mail: Manankohli14@gmail.com

M. Ahmed
e-mail: Muqeem.ahmed@gmail.com

J. Manhas
Department of CS & IT (BC), University of Jammu, Jammu, India
e-mail: Manhas.jatinder@gmail.com

© Springer Nature Singapore Pte Ltd. 2019

V. S. Rathore et al. (eds.), *Emerging Trends in Expert Applications and Security*, Advances in Intelligent Systems and Computing 841,
https://doi.org/10.1007/978-981-13-2285-3_85

drksujatha23@gmail.com

in it, however unregistered individual can just read them. Registered users get to twitter through its site interface, Short Message Service (SMS), or cell phone application software [2].

Essentially, Twitter is a person-to-person communication site. Jack Dorsey and his associates developed it in March 2006. With its launch, it began picking up fame and now it winds up as one of the world's most renowned long-range interpersonal communication sites.

Twitter is particularly an interesting platform because of its concept of hashtags (#) and sentiment analysis. In this paper, we are mostly going to deal with sentiment analysis particularly political sentiments of user with respect to their favourite political parties or candidates. In this research paper, we will try to find the answer to two research questions namely RQ1 and RQ2.

RQ1. Can Twitter data predict the election outcomes?
RQ2. What are the shortcomings of these predictions?

We will discuss these two problems in this paper. First, we will discuss previous research in this field, and then the techniques and models used until now. Second, we will compare various tools and models used in these researches. Then try to find out merits and demerits of previous research. Finally, we will conclude our work and come up with better future suggestions.

2 Related Work

O'Connor et al. [3] measured sentiments of people preconceived idea derived from polls conducted on the renowned and famed site Twitter. Researcher collected data from Twitter API that demonstrates the consumer confidence and political opinion and analyzes movements in the polls. After classification and analyzing the results, it shows that this method predicts public opinion about polls with great accuracy rate hence from its potential it reveals that this technique be adopted as a supplement with traditional polls methods. This research work is done in the USA where the literacy rate is 99% and more than 90% are internet users [4].

Conover et al. [4] developed a method for foreboding the political alignment of active Twitter members for 2010 U.S. midterm elections. They use Support Vector Machine for classification of data, with semantic analysis to identify hidden information of vast data generated by the different user. For classification and communication, network-clustering algorithm is used. After mining, it was revealed that this method can be used for prediction of election analysis with (91%) accuracy rate [3].

Bermingham and Smeaton [5] used Twitter data in predicting the Irish general election 2011 for five main parties. They performed supervised learning for sentiment analysis and Adaboost classifier for classification. Researchers implemented Adaboost MNB classifier for classification, which accomplishes 65.09%

drksujatha23@gmail.com

classification accuracy in their analysis work After mining, they observed that a party namely Fine Gael (FG) has a more winning chance in contrast to others rivalry parties, with Mean Absolute Error 3.67% [4].

Tariq et al. [6] used Twitter data in the prediction of election outcomes. Through the website Twimemachine, data is collected. Researchers implemented various classification algorithms like Naïve Bayes, Support Vector machine, and CHAID decision tree in election prediction with the Rapid Minor tool in the mining of metadata obtained from tweets, respectively. For their research work, they choose three main renowned political parties namely Pakistan Muslim League Nawaz, Pakistan Tehreek e Insaaf, and Muttahida Qaumi Movement. After scrutinizing of data set investigator cogitated that Pakistan Tehreek-e-insaaf would win the election but the actual winner was Pakistan Muslim League Nawaz. The main drawback of this work was that they selected Twitter data for prediction which was used mostly by urban population (that are less than 10%) [5].

Gaurav et al. [7] attempted in analyzing the predictive power of Twitter data for election outcomes of Latin America 2013. Researchers collected approximately 13 billion tweets using Gnipdecahose for Venezuelan, Ecuadorian and Paraguayan Presidents election. Finally, the boffins concluded that by applying Moving Average Aggregate Probability (MAAP) one can accurately predict the election outcomes with less error rate (0.03%) [6].

Wani and Alone [8] analyzed the impact of social media on heterogeneous Indian politically system. In India, 65% populations are below 35 years of age and use social media as a platform to share their views, ideas, communication, sentiments, etc., and these factors play a key role in analyzing election process. The researchers use Twitter data in surveying the user opinion about political parties' and extraction of actual relevant data is done using topic modelling techniques with KNN algorithm. After proper analysis, they came out with results that social media has a great influence on the Indian population. They also found that BJP, which emerged as the winning party in 2014 election, had strategically used social media for their campaign. The main drawback of this paper is that researcher used only Twitter data for election prediction of a vast and heterogeneous country like India, where only small portion of peoples used Twitter less than (15%) [7].

Song et al. [9] the main motive of this research work is to predict the presidential election of a country namely Korea for the year 2012 using Twitter data. The extraction, processing, and analysis of vast twitter data carried out using various data mining techniques like, multinomial topic modelling and network analysis. After analysis, it was revealed that this technique can be used to retrieve fruit-full social trends and hybrid networks content generated through Twitter during election [10].

Almatrafi et al. [11] in their paper attempted to build a model on location base sentimental analysis using tweets for election outcomes. Their main purpose was to collect tweets from a different location for studying human interest and make a

drksujatha23@gmail.com

decision. Researchers collected data using Twitter API tools and Naïve Bayes classification algorithm for examination of two main parties namely AAP and BJP, respectively. After proper examination, it reveals that AAP has a strong network in Delhi but BJP is far ahead from AAP in rest of India. This work clearly depicts that Delhi has a large number of Twitter users as compared to the other state of India. This is because people residing in Delhi has more access to Internet compared to other states [9].

Khatua et al. [12] discussed the idea of Twitter data in predicting election results of large and politically diversified country India. The main challenge was collection of the vast amount of hybrid data because of the prevalence of mainstream, as well as local areas parties, participated in the general election. Lexicon method for sentimental analysis and OLS Regression model for vote swing with Mean Absolute Error for error detection was used [11].

Srivastava and Jain et al. [13] the main aim of this research work was to use Twitter data in predicting the 2015 Delhi Assembly election. For their analysis, they choose three main parties namely AAP, BJP, and Congress, respectively. The Twitter data set was collected using Twitter search engine API, and then various classifier used like Naïve Bayes and Support Vector Machine for classifying the tweets. After proper examination of the tweets, they concluded that AAP had more winning chance as compared to other rivals parties. This model predicted accuracy up to 93.7% with RMSE 6.2% [12].

Wicaksono et al. [14] proposed a politically forecasting model using social media for US presidential election 2016. The collection of diverse data its implementation and processing were the main phases involved in this research work. Researchers did data collection through Twitter search API. Data implementation and processing being done out using advanced Naïve Bayes classification algorithm. After examination, the researcher prediction model brings out Hillary Clinton as a winner candidate but in reality, Donald Trump emerged the winner. This is because one Twitter user may have multiple accounts, which allow him to post more than one tweet favouring his candidate or party. Nevertheless, in election process, one voter can cast only one vote. It means data set obtained proved ambiguous [13].

Safiullah et al. [15] investigated the predicting power of social media especially Twitter in election outcomes. They use Regression Analysis techniques for data analysis with Root Mean Square Error. After examining dataset properly, it was revealed that social media buzz especially Twitter can be used as a healthy indicator in election prediction [14].

Yaqub et al. [1] tried to understand the diverse type of political contents during 2016 US election using twitter data. The unit of analysis was sentiments and retweets of peoples made before the election. Finally, they concluded that user sentiments in form of tweets were mainly negative, as compared to positive and neutral tweets for both the candidates of US presidential election 2016 [15].

drksujatha23@gmail.com

3 Findings of the Literature Survey

In this area, we will talk about different strategies/tools utilized by various authors in their research works. From the above literature review, it is concluded that most of the researchers used classified approach. For this, they used a classification algorithm like Support Vector Machine (SVM), Naïve Bayes, and Decision Tree. For error detection, they used Mean Absolute Error (MAE). Researchers studied different locations, including countries and cities. Each city had different characteristics hence varied parameters such as like Literacy Rate, Internet Literacy, etc. After proper analysis of above-said literature and techniques, we can conclude that:

1. One can predict the election outcome with Twitter Data only in developed countries like Germany, France, USA and Italy. Because in these nations, the literacy rate is above 99% and more than 80% population access Internet.
2. One cannot predict the election results with Twitter Data in developing nations like India Pakistan, Srilanka, etc. Because in these nations average literacy rate is below 70% and only small portion of the population uses the Internet (especially Twitter less than 10%).
3. Twitter result may generate wrong information as one person may have more than one Twitter account. Therefore, such user can post more than one tweets from different accounts, which may generate vague information.
4. In developing nations, it may produce biased results because mostly it used in urban areas, which consist of a small portion of the population (less than 30% in case of India and Pakistan).
5. There is not a normally acknowledged method for "counting votes" on Twitter. Current research has used exponentially volume of tweets, randomly selected users, and varied flavours of sentiment analysis.
6. Data obtained from twitter is not reliable because if we do a random sample of likely voters, Twitter user who have decided to express their opinion only considered for analysis. Moreover, rest Twitter users excluded from forecasting, which may lead ambiguous and erroneous conclusion.

From above, it is clear that we cannot find a single research paper that has universally accepted technologies or models that can accurately predict the election outcomes. The main purpose of this research work is to answer two questions. So our answer to Q No 1 is No. As for the answer to Q No 2, we have found some limitations, which are as below:

- Limited population use it
- Literacy Rate of Nations
- Actual user of Twitter
- Only elite class use it
- Retweets are considered for predictions
- Mostly used in urban part of countries (like India, Pakistan and Bangladesh)
- Non-restriction on rumours and fake news.

drksujatha23@gmail.com

4 Conclusion and Future Work

From above survey, one can conclude that the predictive power of Twitter with respect to elections has been extraordinarily overstated, and needs a great deal of refinement and up gradation in future. The information obtained from Twitter only considered as an important indicator for election prediction. Which until now sufficiently and accurately does not predict the election outcomes. Keeping in mind the end goal to anticipate the election results not only Twitter data as well as all other web-based social networking (Like Facebook, YouTube, and Google +) online and offline component of media should keep into consideration. Data obtained from every one of these sources should mine appropriately by experts. Then some advanced and hybrid tools and techniques like Python, R-programming language, Hadoop, Cassandra, Plotly, and Rapid-miner should apply in getting useful insights.

Further, an all-inclusive level method ought to be proposed or develop, which can precisely foresee the election results. Additionally, for an exact forecast of election result just from Twitter, it must be important that the Twitter specialist evacuates counterfeit records. It must enlist clients simply after their appropriate confirmation through personality archives, for example, Driving License, Passport, and so forth.

References

1. Yaqub U, Chun SA, Atluri V, Vaidya J (2017) Sentiment based analysis of tweets during the US presidential elections. In: Proceedings of 18th annual international conference on digital government research—dg.o '17, pp 1–10
2. Tweeting Made Easier (2017) Accessed 7 Nov 2017
3. O'Connor B, Balasubramanyan R, Routledge BR, Smith NA (2010) From tweets to polls: Linking text sentiment to public opinion time series, From tweets to polls Link, May, pp 122–129
4. Conover MD, Gonçalves B, Ratkiewicz J, Flammini A, Menczer F (2011) Predicting the political alignment of Twitter users. In: Proceedings—2011 IEEE international conference on privacy, security risk Trust IEEE international conference on social computing PASSAT/SocialCom 2011, pp 192–199
5. Bermingham A, Smeaton AF (2011) On using Twitter to monitor political sentiment and predict election results. Psychology 2–10
6. Mahmood T, Iqbal T, Amin F, Lohanna W, Mustafa A (2013) Mining Twitter big data to predict 2013 Pakistan election winner. In: Multi Topic Conference INMIC 2013 16th International, pp 49–54
7. Gaurav M, Srivastava A, Kumar A, Miller S (2013) Leveraging candidate popularity on Twitter to predict election outcome. In: Proceedings of 7th workshop on social network mining and analysis—SNAKDD '13, pp 1–8
8. Wani G, Alone N (2014) A survey on impact of social media on election system. Int J Comput Sci Inf Technol 5(6):7363–7366
9. Song M, Kim MC, Jeong YK (2014) Analyzing the political landscape of 2012 Korean presidential election in Twitter. IEEE Intell Syst 29(2):18–26

drksujatha23@gmail.com

10. Lab CM et al (2014) Centre for culture, media & governance on media framing of 2014 Indian election. Idea 4(2):58
11. Almatrafi O, Parack S, Chavan B (2015) Application of location-based sentiment analysis using Twitter for identifying trends towards Indian general elections 2014. In: Proceedings of 9th International Conference on Ubiquitous Information Management and Communication—IMCOM '15, pp 1–5 (2015)
12. Khatua A, Khatua A, Ghosh K, Chaki N, Can #Twitter-Trends predict election results? Evidence from 2014 Indian general election. Proceedigns of Annual. Hawaii International Conference on System Sciences, vol 2015, pp 1676–1685, Mar 2015
13. Srivastava R, Bhatia MPS, Kumar H, Jain S (2015) Analyzing Delhi assembly election 2015 using textual content of social network. ACM International Conference Proceeding Series, vol 25–27, pp 78–85, Sept 2015
14. Wicaksono AJ, Suyoto, Pranowo (2017) A proposed method for predicting US presidential election by analyzing sentiment in social media. In: Proceeding—2016 2nd International Conference on Science in Information Technology ICSITech 2016 Information Science Green Society Environment, pp 276–280
15. Safiullah M, Pathak P, Singh S, Anshul A (2017) Social media as an upcoming tool for political marketing effectiveness. Asia Pacific Manag Rev 22(1):10–15

drksujatha23@gmail.com

An Intelligent Framework for Sentiment Analysis of Text and Emotions—Implementation of Restaurants' Case Study



Esha and Arvind Kumar Sharma

Abstract Nowadays, online purchasing of products is becoming very popular. The thing which counts is that whether the purchased product satisfies the need of a consumer or not. The buyer does not have clear idea about the product initially, due to the lack of straight deal or physical touch of the product. So, to solve the issue of reliability, consumer moves to review analysis available online regarding the product. Hereto the limitation is that a single review may not be trustworthy due to fake or false review about the particular product. To overcome such issues, a certain satisfying model based on online product ratings needs to be worked out which could incorporate the available reviews. The primary aim of this research paper is to implement an intelligent framework for sentiment analysis of text and emotions and apply machine learning approach for computing the effectiveness and efficiency of overall ratings by the consumer for particular item or product.

Keywords Sentimental analysis · Machine learning algorithms
Framework · Emotions

1 Introduction

To avoid crowded markets and the time-consuming factors along with un-comforts, people prefer e-shopping. In this model, consumer directly buys product/service from the sellers without any intermediator through the Internet. Just by comfortably sitting before computer, consumers visit various websites or e-stores providing 24 × 7 h of service and connected with Internet from their place itself. This provides a convenient environment to shop online and availing services. Another inviting feature is that even on holiday, consumers are free to shop online and

Esha · A. K. Sharma (✉)
Career Point University, Kota, India
e-mail: drarvindkumarsharma@gmail.com

Esha
e-mail: eshatyagi26@gmail.com

© Springer Nature Singapore Pte Ltd. 2019

V. S. Rathore et al. (eds.), *Emerging Trends in Expert Applications and Security*,
Advances in Intelligent Systems and Computing 841,
https://doi.org/10.1007/978-981-13-2285-3_86

drksujatha23@gmail.com

availing services [1] without waiting in a long queue and searching for a particular product in the store. As a variety of products can be having on online stores, consumers feel difficult in the selection of the products and to take decision regarding the item. This inconvenience for the consumers is considered by the researchers and various data regarding sentiments or emotions is collected as per their priority for products or services for mining. Due to WWW, use of the Internet has rapidly increased among the people. Now the problem is to make a proper decision using huge information available on the social sites. Considering the priority of online purchase of the products or services, people have to trust reviews about products available on WWW. Based on reviews and ratings, the decision is taken by the consumers to purchase unique product or to avail service. The framework that recommends the consumer about the appropriateness of product/service can be termed as recommender system [2]. Useful results are also available for consumers to support with the large size of information on social network. People are becoming both an information producer and consumer due to increase in Web which assists them to solve risky problems. By the evolved framework decision-making process will become easier regarding information available via online textual reviews. Reviews and reviewers are taken as helper for choosing product/service. Large information about product given by users or consumers is also used for reviews or opinions. Target-advertising, personal-marketing, and information-retrieval are used widely by recommender system that is significant part in e-commerce application [3].

2 Sentiment Analysis

Sentimental analysis process may be done with help of set of interrelated subprocesses. The complete analysis provides a structure for natural language text obtained from the compound and threatening unstructured text. Furthermore, on basis of some problems involved, sentiment analysis is done using a common framework with the existing system combination [4]. The increasing use of the internet has increased an immense data on social sites. This information is available in posts, news, articles, comments, reviews, etc. Reviews, comments, and opinions of users play a vital role in popularizing any product or service or bringing their graph down. Much depends on the response based on specific presentation of product or service. Data consisting of such reviews or opinions have an immense potentiality to be analyzed. Buying and selling of products/services with transition information via social network is carried out in e-commerce. Several reviews for a specific product play a vital role in mining the opinion of people. It is a timely need to design models or framework to automate the classification of distinctive reviews. The basis of Sentimental Analysis Sentiment analysis is used to identify attitude or opinion or emotion expressed by people towards products/services. Sentiment

drksujatha23@gmail.com

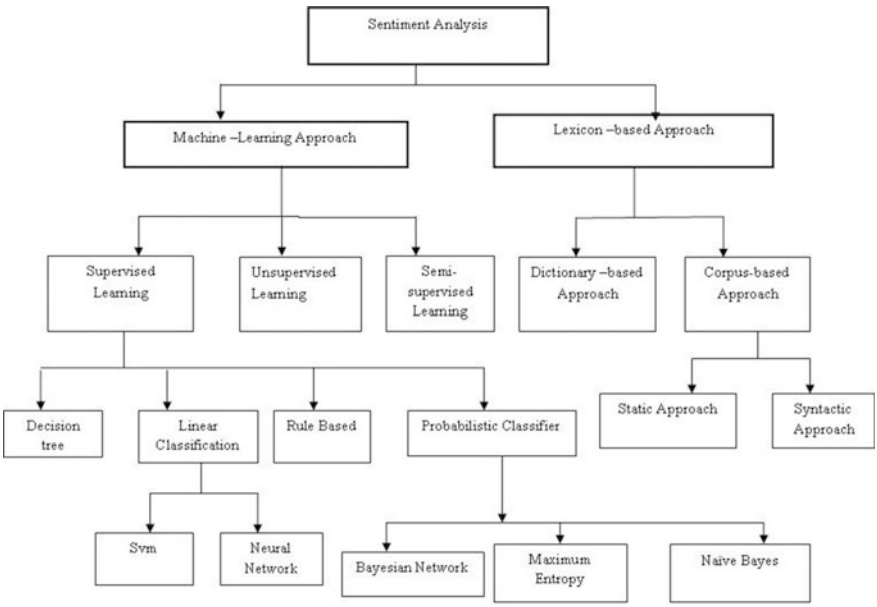


Fig. 1 Sentiment analysis approaches [8]

analysis, referred to as sentiment mining speaks to as utilization of regular dialect handling, content examination and computational etymology to research and concentrate subjective data from source materials. Sentiment analysis approaches are classified into the following categories as shown in Fig. 1.

3 Research Methodology

The existing systems of research methodology used for sentiment analysis of emotions and rating based prediction of consumer's need might be elaborated.

3.1 Features of Existing Research Methodology

Review level, sentence level, and phrase level are the levels used to perform sentiment analysis. Review of sentiment is classified into the review-level analysis and sentence-level analysis towards prevailing polarities, i.e., positive/negative/neutral.

The specific feature of product is used to perform mining for the expression of user and attitude when the analysis is done at the phrase level. Self-supervised and lexicon-based technique recommended by Zhang et al. [5] is used to find polarity of

drksujatha23@gmail.com

reviews containing both text and emotion within it. To find both new aspects and relevant depending on concept of experts, Recommender System proposed by Lee et al. [6] experts have recommended user based popularity to analyze user ratings.

3.2 Drawbacks of Existing Research Methodology

The prevailing system is based upon binary sentiment, i.e., positive/negative. Users could classify in the prevailing system but the mining of user's sentiment is not done.

The methodology contains the following key indicate recommended framework to be developed: In the present model, a method for sentiment-based rating prediction is proposed. Rating of user's emotions is to be done in our work. To begin with the review of user's sentiments is mined. Thereafter the main features of product are described by finding out the sentimental words.

While considering Items or products sentiment of unique user is computed through leverage sentiment dictionaries. For mining process, the sentimental words from the reviews provided by the user are used. Rate prediction for sentiment is used. User interest preferences are focused in parallel. The reason for sentiment spreads among the trusted users is projected by user sentiment influence.

User sentiment similarity, interpersonal sentimental influence and items reputation similarity are three factors joined upon probabilistic matrix factorization model for an accurate recommendation. User's emotions are given. The result based on an experiment where the performance of rate prediction improvement is the key factor. The methodology contains the following key points: The methodology discusses simulating Machine Learning Algorithms in Sentiment Analysis and to study relationships among online reviews for smart phone products and the revenue of performance.

4 Implementation

4.1 Modules of Framework

In this part, the various modules are developed for the intelligent framework to analyze emotions. Dataset feedback and user's rating based is unique features of our developed framework and separated here using novel approach. From social user's ratings and reviews valuable trends are identified by our recommended framework. The framework is also used to identify the aspect of social user's sentiments. On the basis of user's reviews to extract reliable models is also a unique feature of our research work. Finally, item's reputation, interpersonal sentiment influence, and user sentiment commonality are the three parameters tested by the recommended model.

drksujatha23@gmail.com

The modules to analyze sentiments or emotions of the consumers to get accurate recommendation are—Data Preprocessing for Dataset, Extracting Product Features, User Sentimental Measurement, Sentiment Evaluation, and Ratings Are Sentiments (RAS) Module 4.2 Modules Description Data Preprocessing for Dataset The initial module is designed for data preprocessing phase for the collected dataset from the http://www.yelp.com [7] web source rating data set. Loading dataset as input in this module is a major part of framework. Product items dataset, user ratings dataset, and user feedback dataset are used in this framework very effectively. Extracting of Product Feature This framework has a unique feature to extract specific features of products and services for textual reviews.

From the named entities list of product feature is obtained with few attributes of products, item, service, etc.

The relationship of topics, words, and reviews are the models which are utilized using SVM model. User's review is looking upon initially as group of words with order considering in it to construct the vocabulary. After that stop words, noise words and words, with its sentiment degree and contrary words are filtered.

4.2 Experimental Work

This section presents the experimental analysis of our system. Simulation of research work and its usage is as under (Figs. 2, 3, 4, 5, 6 and 7).

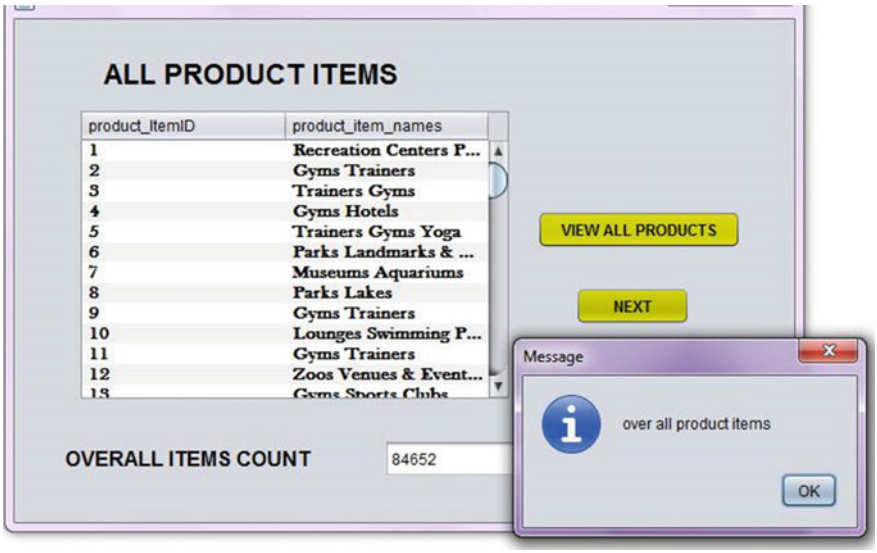


Fig. 2 Framework shows list of product items observed by the user

drksujatha23@gmail.com

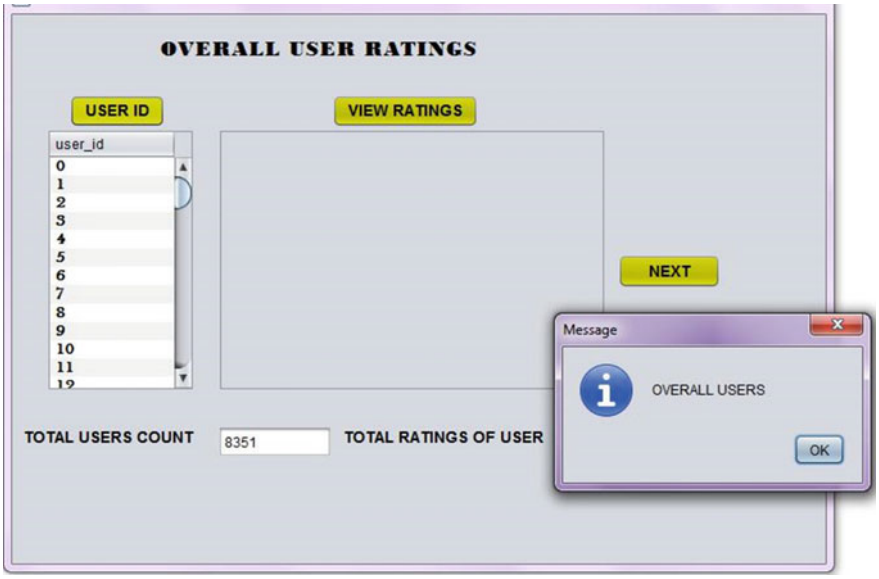


Fig. 3 Window shows overall users involved in rating the items in sentiment analysis

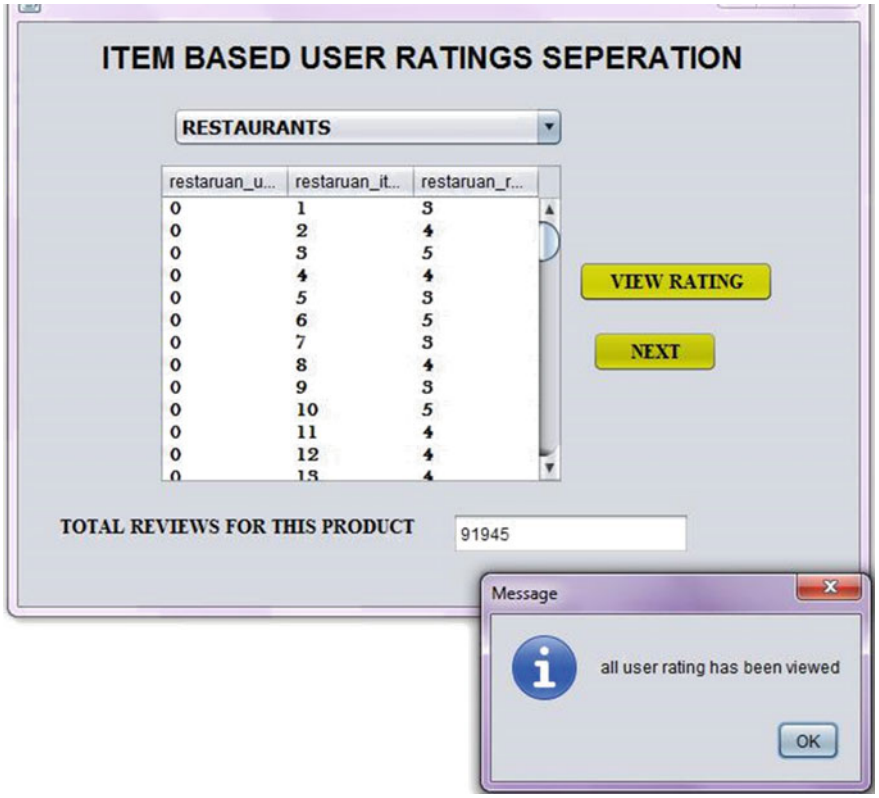


Fig. 4 Window shows item based user ratings separation

drksujatha23@gmail.com

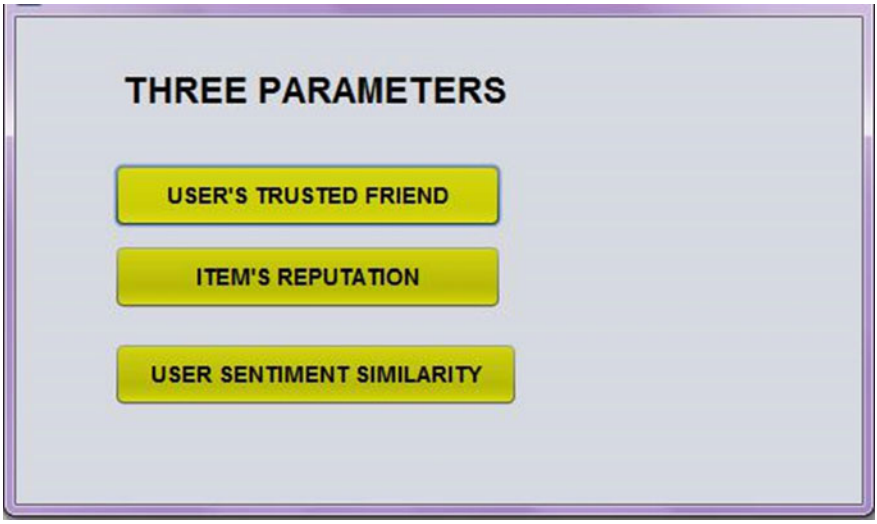


Fig. 5 Proposed three parameters involved in accurate user ratings

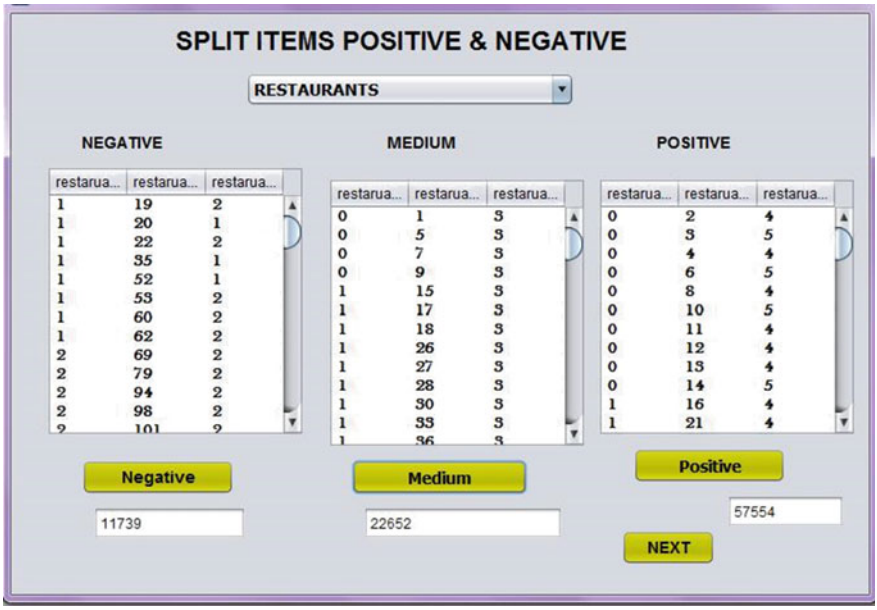


Fig. 6 Calculate total negative, medium, positive ratings

drksujatha23@gmail.com

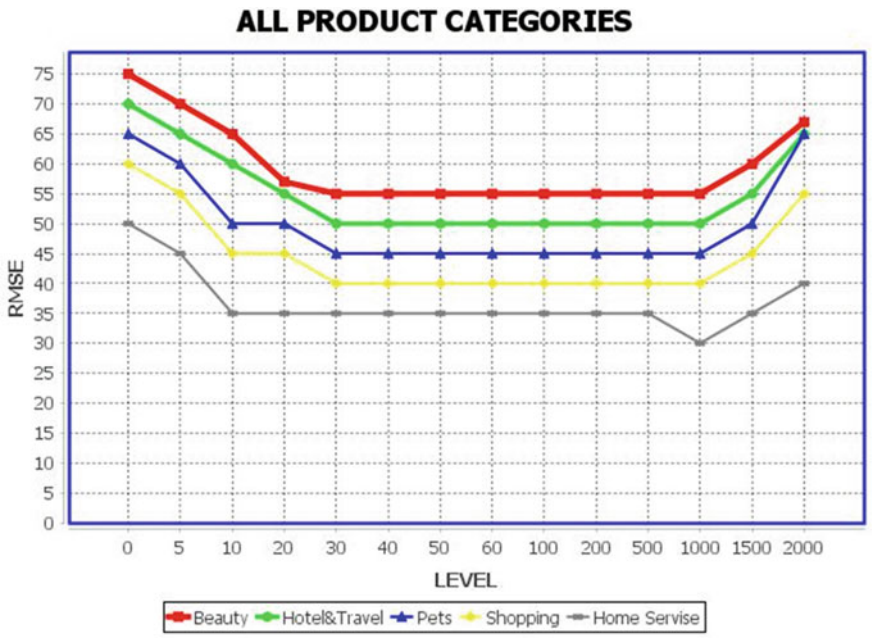


Fig. 7 Sentimental similarity feedback for all product categories

Table 1 Performance of presented framework using different ML algorithms

Algorithm used	Accuracy obtained	Precision	Recall	F-measure
SVM	0.823	0.836	0.835	0.834
Nave Bayes	0.731	0.675	0.690	0.680
Forest	0.812	0.768	0.769	0.767
Tree	0.334	0.160	0.600	0.330

The presented framework has been tested using different machine learning algorithms with TF-IDF, using the dataset acquired from the yelp online. The dataset is divided into two equal parts, i.e., training and testing data. The performance of SVM algorithm for the recommended framework is highly appreciable and exceptional. Table 1 shows the performance of the presented framework using different machine learning algorithms.

In the above table, the three performance evaluation metrics of machine learning algorithms are utilized as—Precision, Recall, and F-measure. On the basis of the given parameters, the accuracy is calculated on the basis of users' reviews or users' sentiments. With the help of obtained results, it concludes that our recommended model accuracy is more than the existing model (Fig. 8).

drksujatha23@gmail.com

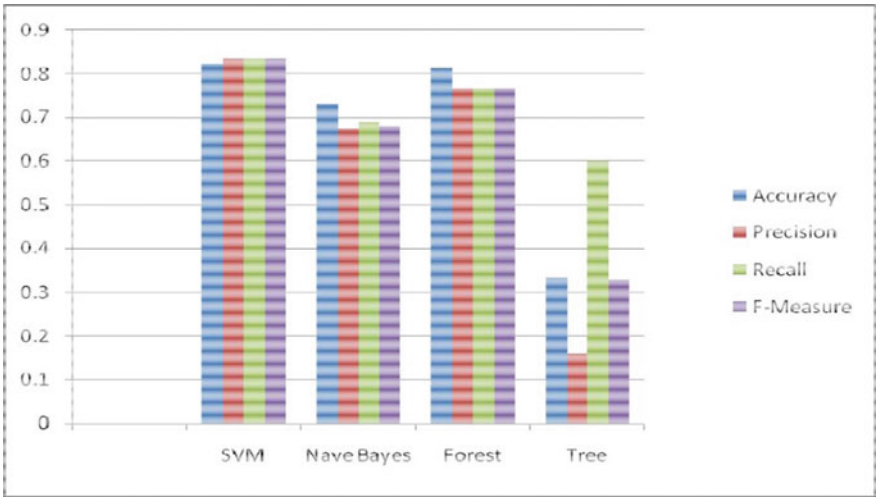


Fig. 8 Performance analyses of different ML algorithms

5 Conclusion

Social user's textual review has gained much more importance these days. The basic purpose of sentiment analysis is to predict polarity of sentiments and emotions reflected in a given sentence. To identify the true nature of the expressed sentiments is a need of time, i.e., they reflect positive/negative feeling about the particular topic. After reviewing the sentiments of social users, sentiment extraction is done using a novel framework in this paper. User sentiment's similarity, interpersonal sentiment influence, item reputation similarity, etc. are done with matrix factorization in the developed framework. The rate of prediction achieved in exact to represent user's preferences. Interpersonal sentiment influence and relationship developed with users and friends from sentiment angle is also explored. User's sentiment has been measured quantitatively, from the user's items leverage to obtain item's reputation. Three sentiment parameters are demonstrated as result of our simulation which provides a great support to rate the prediction of users. Besides it, a major improvement is the identification of real-world dataset based approach and presented as the Sentiment Analysis of Emotions and Rating Based Consumer Sentiments. Testing of the research model by using different ML algorithms on datasets collected of social user's reviews or sentiments from YELP to find polarity of user's sentiments and texts is also done to know whether they are positive or negative in relation to three words related to sentiment of products. The performance resulting models tested the value of accuracy, precision, recall, and f-measure of different ML algorithms.

Finally, the SVM algorithm achieving higher accuracy, so, it concludes that SVM is robust and better one. Experimental analysis also shows the efficiency of

drksujatha23@gmail.com

the proposed framework. In future work, this framework could be extended and certain objectives can be fulfilled in this system. First, the system instead of taking the textual reviews can also consider audio reviews providing an extra facility for social users. Developing an app with these features is a broad parameter. The datasets can be increased further to get all sort of polarity and features of words. Besides it, the spell checking and the nearby word approach could also be considered.

References

1. Salakhutdinov R, Mnih A (2008) Probabilistic matrix factorization. In: NIPS
2. Yang X, Steck H, Liu Y (2012) Circle-based recommendation in online social networks. In: Proceedings of the 18th ACM SIGKDD international conference on KDD, New York, NY, USA, Aug 2012, pp 1267–1275
3. Jiang M, Cui P, Liu R, Yang Q, Wang F, Zhu W, Yang S (2012) Social contextual recommendation. In: Proceedings of the 21st ACM international CIKM, pp 45–54
4. Ganu G, Elhadad N, Marian A (2009) Beyond the stars: improving rating predictions using review text content. In: 12th international workshop on the web and databases (WebDB 2009), pp 1–6
5. Zhang W, Ding G, Chen L, Li C, Zhang C (2013) Generating virtual ratings from Chinese reviews to augment online recommendations. ACM TIST 4(1):1–17
6. Pang B, Lee L (2008) Opinion mining and sentiment analysis. ACM J Found Trends Inf Retr Arch 2(1–2)
7. [Online] http://www.yelp.com
8. Kumari U et al (2017) A cognitive study of sentiment analysis techniques and tools: a survey. IJCST 8(1)

drksujatha23@gmail.com

Author Index

A
Advani, Nilesh, 393
Agarwal, Niket, 425
Agarwal, Swati, 97
Agrawal, Apoorva, 211
Agrawal, Saroj, 11
Ahmed, Muqeem, 725
Akankasha, 477
Arora, Bhavna, 477, 505
Arora, Yojna, 55
Arya, Nancy, 111
Avinash, Sharma, 1

B
Baid, Palak, 307
Bandil, Devesh, 357
Banyal, R. K., 617
Barddhan, Alok, 521
Baskaran, K., 719
Bhagat, Neeraj, 505
Bhardwaj, Shweta, 63, 73, 513
Bharti, Neha, 633
Bhat, Anjum Zameer, 691
Bhatti, Dharmendra G., 529
Bhargava, Divyang, 211
Bharodiya, Anil K., 47
Bhavani, Nallamilli P. G., 41
Borgaonkar, Pranjali, 625

C
Chandna, V. K., 675
Chande, Swati V., 385
Chaplot, Neelam, 307
Chaturvedi, Poonam, 117
Chaudhary, Sunita, 111
Chitra Rajagopal, P., 565, 575
Choudhary, Kirti, 633
Choudhary, Renu, 133
Choudhury, Tanupriya, 433, 555, 565, 575, 583, 593

D
Dadhich, Mamta, 461
Dalal, Surjeet, 327
Degadwala, Sheshang, 189
Desai, Hiral, 139, 379
Dhaked, Devender, 11
Dhingra, Madhavi, 125
Dhingra, Prince, 433
Diwan, Bahul, 513
Dixit, Abhishek, 337, 535
Dixit, Aniket, 211

E
Esha, 733

F
Fr. Augustine George, 699

G
Garg, Hemant Kumar, 683
Gaur, Sanjay, 189, 219, 379
Gautam, Diwakar, 219
Gautam, Geetika, 633
Gautam, Jyoti, 245
Gianey, Hemant, 425
Gill, Sumeet, 167
Goel, Shivanshi, 583
Gonsai, Atul M., 47, 393
Goyal, Dinesh, 55

© Springer Nature Singapore Pte Ltd. 2019

V. S. Rathore et al. (eds.), *Emerging Trends in Expert Applications and Security*,
Advances in Intelligent Systems and Computing 841,
https://doi.org/10.1007/978-981-13-2285-3

drksujatha23@gmail.com

Goyal, Hemlata, 177
Goyal, Jayanti, 365
Goyal, Rajeev, 485
Goyal, Sandip Kumar, 1
Gupta, Ankita, 521
Gupta, Arushi, 245
Gupta, Gaurav, 675
Gupta, Kavya, 245
Gupta, Priya, 229
Gupta, Shashank, 425
Gupta, Shloak, 147

H
Hanumanthappa, M., 253, 699

I
Indumathi, R., 273

J
Jadon, Priyanshu, 659
Jadon, Rakesh Singh, 125
Jahan, Wani Shah, 81
Jain, Nidhi, 521
Jain, S. C., 125
Jalem, Raj Srujan, 493
Jat, Subhash Chandra, 239
Jayalakshmi, D, 261
Jeba Nega Cheltha, C., 541
Jha, Rajan Kumar, 541
Jindal, Upasna, 327
Joshi, Nisheeth, 177

K
Kamal, Kumar, 1
Kasiviswanath, N., 707
Keswani, Bright, 291
Khan, Saim, 63
Khandelwal, Ayush, 211
Khandelwal, Sarika, 219
Khanna, Shaurya, 73
Khatoon, Arfiha, 97
Khurana, Anirudh, 73
Koli, Abdul Manan, 725
Kotwal, Shallu, 469
Kumar, Ajay, 605
Kumar, Praveen, 451, 521, 575
Kumar, Sarvesh, 97
Kumar, Vijay, 315
Kumar, Vikas, 555
Kumari, Archana, 565
Kumbharana, C. K., 401, 409

L
Lad, Kalpesh, 299
Lakhwani, Kamlesh, 425
Lalwani, Paras, 451
Lamba, C. S., 203, 239
Lee, JangYeol, 547

M
Mahajan, Arpana, 189
Maheshwari, Shikha, 23, 211
Majumdar, Rana, 445
Manhas, Jatinder, 469, 725
Mathur, Richa, 357
Mathuria, Manish, 11
Meenakshi, 167
Mehta, Darshan M., 529
Mehta, Inderpal Singh, 583
Mishra, Durgesh Kumar, 659
Mishra, Shweta, 555

N
Nagalavi, Deepa, 253
Naidu, Vikas Rao, 691
Nam, ChoonSung, 547
Naveena Shri, P. C., 719
Nazir, Nahida, 651

P
Padhya, Dipal, 103
Pant, Pooja, 433
Pareta, Chetana, 605
Park, JinHyuck, 547
Pathak, Vibhakar, 357
Patel, Bankim, 319, 347, 667
Patel, Himadri, 667
Pawar, Kriti, 493
Ponmagal, R.S., 41
Poornima, E., 707

R
Raghavan, R., 273
Raj, Ankur, 229
Rajab, Sharifa, 643, 651
Rajab, Waheeda, 643
Ram, Shrwan, 147
Ranpara, Ripal, 401, 409
Rathod, Chetan, 393
Rathore, Jyoti, 291
Rathore, Karan, 451
Rathore, Vijay Singh, 23, 203, 211, 239, 291, 461, 535

drksujatha23@gmail.com

Rawat, Seema, 451
Rishi, O. P., 197
Roshan, Rakesh, 197

S
Sachin, Sharma, 1
Sai Sabitha, A., 433, 555, 583, 593
Sankar, S., 261
Saravanan, G., 41
Saxena, Akash, 365
Saxena, Khushboo, 365
Saxena, Mohit, 117
Saxena, Prashant S., 23
Sehgal, Anchal, 337, 535
Sharma, Amit, 469
Sharma, Anukrati, 197
Sharma, Archana, 575
Sharma, Arvind, 617
Sharma, Arvind Kumar, 605, 625, 733
Sharma, Avinash, 315
Sharma, Bhavna, 117
Sharma, Chilka, 177
Sharma, Dakshita, 593
Sharma, Harish, 625
Sharma, Manu, 89
Sharma, Nirmala, 625
Sharma, Pankaj Kumar, 133
Sharma, S. K., 379
Sharma, Sandeep Kumar, 203
Sharma, Sanjiv, 485
Sharma, Satyendra K., 139
Sharma, Shilpi, 157
Sharma, Vinod, 643, 651
Sharma, Vipul, 315
Sharma, Viverdhana, 469
Sheik, Alam, 379
Sheikh, Mohammed Firdos Alam, 139, 379
Sheth, Jikitsha, 103
Shin, DongRyeol, 547
Shoba Bindu, C., 707
Shrivastava, Pallavi, 31
Singh, Abhishek, 157
Singh, Baldev, 691
Singh, Chetna, 445
Singh, Dhirender, 617
Singh, Harry, 445
Singh, Shashikant, 229
Singh, Sonali, 97
Solanki, Neha, 219
Soni, Diwanshu, 211
Soni, Hemlata, 675
Srivastava, Ashish, 31
Sujatha, K., 41
Sukumar, Shivashankar, 273

T
Tailor, Chetana, 319
Tailor, Jaishree, 299
Taruna, S., 111
Thakor, Devendra, 347
Tiwari, Mahima, 245
Tiwari, Preeti, 385
Tiwari, Vivek, 493
Tripathi, Pragati, 417

U
Updhyay, Arvind K., 485
Urooj, Shabana, 417

V
Venkateshkumar, M., 261, 273
Vijay, Aman, 211
Vijaivargia, Shilpa, 683
Vora, Mital, 409

Y
Yaadav, N. S., 605
Yadav, Seema, 229
Yadav, Surendra, 11
Yusufzai, Asifkhan, 401, 409

drksujatha23@gmail.com