

Asset Analytics

Performance and Safety Management

Series Editors: Ajit Kumar Verma · P. K. Kapur · Uday Kumar

Millie Pant

Tarun K. Sharma

Sebastián Basterrech

Chitresh Banerjee *Editors*

Performance Management of Integrated Systems and its Applications in Software Engineering

 Springer

Asset Analytics

Performance and Safety Management

Series Editors

Ajit Kumar Verma, Western Norway University of Applied Sciences, Haugesund, Rogaland Fylke, Norway

P. K. Kapur, Centre for Interdisciplinary Research, Amity University, Noida, India

Uday Kumar, Division of Operation and Maintenance Engineering, Luleå University of Technology, Luleå, Sweden

The main aim of this book series is to provide a floor for researchers, industries, asset managers, government policy makers and infrastructure operators to cooperate and collaborate among themselves to improve the performance and safety of the assets with maximum return on assets and improved utilization for the benefit of society and the environment.

Assets can be defined as any resource that will create value to the business. Assets include physical (railway, road, buildings, industrial etc.), human, and intangible assets (software, data etc.). The scope of the book series will be but not limited to:

- Optimization, modelling and analysis of assets
- Application of RAMS to the system of systems
- Interdisciplinary and multidisciplinary research to deal with sustainability issues
- Application of advanced analytics for improvement of systems
- Application of computational intelligence, IT and software systems for decisions
- Interdisciplinary approach to performance management
- Integrated approach to system efficiency and effectiveness
- Life cycle management of the assets
- Integrated risk, hazard, vulnerability analysis and assurance management
- Adaptability of the systems to the usage and environment
- Integration of data-information-knowledge for decision support
- Production rate enhancement with best practices
- Optimization of renewable and non-renewable energy resources

More information about this series at <http://www.springer.com/series/15776>

Millie Pant · Tarun K. Sharma ·
Sebastián Basterrech · Chitresh Banerjee
Editors

Performance Management of Integrated Systems and its Applications in Software Engineering

 Springer

المنارة للاستشارات

Editors

Millie Pant
Department of Applied Science
and Engineering
Indian Institute of Technology Roorkee
Roorkee, Uttarakhand, India

Sebastián Basterrech
Department of Computer Science
Czech Technical University in Prague
Ostrava, Praha, Czech Republic

Tarun K. Sharma
Amity School of Engineering & Technology
Amity University Rajasthan
Jaipur, Rajasthan, India

Chitresh Banerjee
Amity Institute of Information Technology
Amity University Rajasthan
Jaipur, Rajasthan, India

ISSN 2522-5162

Asset Analytics

ISBN 978-981-13-8252-9

<https://doi.org/10.1007/978-981-13-8253-6>

ISSN 2522-5170 (electronic)

ISBN 978-981-13-8253-6 (eBook)

© Springer Nature Singapore Pte Ltd. 2020

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, expressed 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

Contents

| | |
|---|----|
| Real-Time Distributed Denial-of-Service (DDoS) Attack Detection Using Decision Trees for Server Performance Maintenance | 1 |
| Mrunmayee Khare and Rajvardhan Oak | |
| Cloud Computing: Vulnerability and Threat Indications | 11 |
| Vaishali Singh and S. K. Pandey | |
| Proposed Algorithm for Creation of Misuse Case Modeling Tree During Security Requirements Elicitation Phase to Quantify Security | 21 |
| Ajeet Singh Poonia, C. Banerjee, Arpita Banerjee and S. K. Sharma | |
| Big Data Analytics for Data Quality Improvement to Enhance Evidence-Based Health Care in Developing Countries | 29 |
| Billy Mathias Kalema and Viola Vivian Busobozi | |
| Securing the Data Deduplication to Improve the Performance of Systems in the Cloud Infrastructure | 43 |
| Nishant N. Pachpor and Prakash S. Prasad | |
| Implementation of Collaborative Filtering for Product Recommendation in E-Commerce to Enhance Scalability and Performance | 59 |
| Niti Vishwas, Tapajyoti Deb, Ashim Saha and Lalita Kumari | |
| A Pre-emptive Goal Programming Model for Multi-site Production and Distribution Planning | 71 |
| Gaurav Kumar Badhotiya, Gunjan Soni and M. L. Mittal | |
| An Analysis of Comorbidities' Role in Diabetes Mellitus and Its Data-Intensive Technology-Based Prediction to Reduce Risk and Diagnostic Costs | 83 |
| M. Venkatesh Saravanakumar and M. Sabibullah | |

| | |
|---|-----|
| Interpreting the Objective Outcome of the Proposed Misuse Case Oriented Quality Requirements (MCOQR) Framework Metrics for Security Quantification | 101 |
| Ajeet Singh Poonia, C. Banerjee, Arpita Banerjee and S. K. Sharma | |
| A Comparative Performance Study of Machine Learning Algorithms for Sentiment Analysis of Movie Viewers Using Open Reviews | 107 |
| Dilip Singh Sisodia, Shivangi Bhandari, Nerella Keerthana Reddy and Abinash Pujahari | |
| A Comparative Study on Different Approaches of Road Traffic Optimization Based on Big Data Analytics | 119 |
| Tapajyoti Deb, Niti Vishwas and Ashim Saha | |
| Comparative Study Between Cryptographic and Hybrid Techniques for Implementation of Security in Cloud Computing | 127 |
| Sumit Chaudhary, Foram Suthar and N. K. Joshi | |
| Functional Module Detection in Gene Regulatory Network Associated with Hepatocellular Carcinoma | 137 |
| Sachin Bhatt, Kalpana Singh and Ravins Dohare | |
| Comparative Analysis of Various Techniques of DDoS Attacks for Detection & Prevention and Their Impact in MANET | 151 |
| Neha Singh, Ankur Dumka and Rakesh Sharma | |
| A Comparative Study of Data Mining Tools and Techniques for Business Intelligence | 163 |
| G. S. Ramesh, T. V. Rajini Kanth and D. Vasumathi | |
| Performance Analysis of E-Governance Citizen-Centric Services Through E-Mitra in Rajasthan | 175 |
| Praveen Kumar Sharma and Vijay Singh Rathore | |
| A Critical Study on Disaster Management and Role of ICT in Minimizing Its Impact | 183 |
| Pratibha Choudhary and Rohit Vyas | |
| Development of Arduino-Based Compact Heart Pulse and Body Temperature Monitoring Embedded System for Better Performance | 189 |
| Sandeep Gupta, Akash Talwariya and Pushpendra Singh | |
| Performance Evaluation of Learners for Analyzing the Hotel Customer Sentiments Based on Text Reviews | 199 |
| Dilip Singh Sisodia, Saragadam Nikhil, Gundu Sai Kiran and Hari Shrawgi | |

| | |
|---|-----|
| Proposed Data Structure for Storage of Metrics Values: Misuse Case Oriented Quality Requirements (MCOQR) Framework Perspective | 211 |
| Sunita Choudhary, C. Banerjee, Ajeet Singh Poonia, Arpita Banerjee and S. K. Sharma | |
| Comparative Analysis of Hindi Text Summarization for Multiple Documents by Padding of Ancillary Features | 217 |
| Archana N. Gulati and Sudhir D. Sawarkar | |
| RC-Network and Comm-Network for Improvement of Research Collaboration and Communication Among Delhi University Teachers | 227 |
| Narender Kumar, Sapna Malhotra, Chitra Rajora and Ravins Dohare | |

Editors and Contributors

About the Editors

Dr. Millie Pant is Associate Professor at the Department of Applied Science & Engineering., IIT Roorkee, India. She has published over 180 research papers and has edited a number of edited volumes and conference proceedings published by Springer. She is Associate Editor, Guest Editor and reviewer for various Springer and Inderscience journals and IEEE Transactions. She has served as General Chair, Program Chair, Session and Track Chair at numerous national & international conferences, and has delivered guest lectures at various leading national and international institutions. She has been involved in international collaboration with MIRS Lab, USA; Liverpool Hope University, UK; and Université Paris-EstCréteil Val-de-Marne, Paris, France.

Dr. Tarun K. Sharma is an Associate Professor at Amity University Rajasthan, India. He holds a Ph.D. in Soft Computing from IIT, Roorkee, and has published over 90 research papers. He has served as General Chair, Program Chair, Track Chair in the SoCTA, SoCPros Conference Series. He has edited a number of edited volumes and conference proceedings published by Springer. He is Associate Editor, Guest Editor and reviewer for various Springer and Inderscience journals and IEEE Transactions. He has delivered guest lectures at various leading national and international institutions. He is member of IET, IANEG, CSTA, and MIRS Lab.

Dr. Sebastián Basterrech is an Associate Professor at the Department of Computer Science, Faculty of Electrical Engineering, Czech Technical University, Prague, He has 70+ research publications to his credit. He is Associate Editor, Guest Editor and reviewer Springer and Inderscience journals and IEEE Transactions. He has acted as Program Chair and Technical Chair at numerous national & international conferences, and has made valuable contributions in areas related to quasi-Newton optimization, random neural networks, reservoir computing, neural computation & soft-computing techniques.

Dr. Chitresh Banerjee is an Assistant Professor at Amity University, Rajasthan, India. He has published over 60 research papers and has also worked as Executive Officer on the Board of Studies at The Institute of Chartered Accountants of India, New Delhi. He is member of 15 international societies and associations. Under the Institute-Industry linkage program, he delivers expert lectures on various themes related to IT. He has authored several books, and has acted as Editor, Associate Editor, Guest Editor and reviewer for numerous national and international journals and conference proceedings.

Contributors

Gaurav Kumar Badhotiya Department of Mechanical Engineering, Malaviya National Institute of Technology Jaipur, Jaipur, Rajasthan, India

Arpita Banerjee St. Xavier's College, Jaipur, India

C. Banerjee Amity University Rajasthan, Jaipur, India

Shivangi Bhandari National Institute of Technology Raipur, Raipur, India

Sachin Bhatt Centre for Interdisciplinary Research in Basic Sciences, Jamia Millia Islamia, New Delhi, India

Viola Vivian Busobozi Department of Informatics, Tshwane University of Technology, Pretoria, South Africa

Sumit Chaudhary Uttaranchal University, Dehradun, India

Pratibha Choudhary Government College of Engineering & Technology, Bikaner, India

Sunita Choudhary Government College of Engineering and Technology, Bikaner, India

Tapajyoti Deb National Institute of Technology, Agartala, India

Ravins Dohare Centre for Interdisciplinary Research in Basic Sciences, Jamia Millia Islamia, New Delhi, India

Ankur Dumka Graphic Era, Deemed to be University, Dehradun, India

Archana N. Gulati Department of Computer Engineering, Datta Meghe College of Engineering, Mumbai, Maharashtra, India

Sandeep Gupta Electrical Engineering Department, JECRC University, Jaipur, India

N. K. Joshi Uttaranchal University, Dehradun, India

Billy Mathias Kalema Department of Informatics, Tshwane University of Technology, Pretoria, South Africa

Mrunmayee Khare Department of Computer Engineering, Pune Institute of Computer Technology, Pune, India

Gundu Sai Kiran National Institute of Technology Raipur, Raipur, India

Narender Kumar Department of Mathematics, Gargi College, University of Delhi, New Delhi, India

Lalita Kumari National Institute of Technology, Agartala, India

Sapna Malhotra Department of Mathematics, Gargi College, University of Delhi, New Delhi, India

M. L. Mittal Department of Mechanical Engineering, Malaviya National Institute of Technology Jaipur, Jaipur, Rajasthan, India

Nishant N. Pachpor P.I.E.T, Nagpur, India

Saragadam Nikhil National Institute of Technology Raipur, Raipur, India

Rajvardhan Oak Department of Computer Engineering, Pune Institute of Computer Technology, Pune, India

S. K. Pandey Department of Electronics & Information Technology, Ministry of Communications & IT, Government of India, New Delhi, India

Ajeet Singh Poonia Government College of Engineering and Technology, Bikaner, India

Prakash S. Prasad P.I.E.T, Nagpur, India

Abinash Pujahari National Institute of Technology Raipur, Raipur, India

T. V. Rajini Kanth Department of CSE, SNIST, Hyderabad, India

Chitra Rajora Department of Commerce, Gargi College, University of Delhi, New Delhi, India

G. S. Ramesh Department of CSE, VNR VJIET, Hyderabad, India

Vijay Singh Rathore JECRC, Jaipur, Rajasthan, India

Nerella Keerthana Reddy National Institute of Technology Raipur, Raipur, India

M. Sabibullah PG and Research Department of Computer Science, Jamal Mohamed College (Autonomous), Tiruchirappalli, TN, India

Ashim Saha National Institute of Technology, Agartala, India

Sudhir D. Sawarkar Datta Meghe College of Engineering, Mumbai, Maharashtra, India

Praveen Kumar Sharma Mewar University Chittorgarh, Chittorgarh, Rajasthan, India

Rakesh Sharma Government of Jharkhand in Higher & Technical Education, Jharkhand, India

S. K. Sharma Modern Institute of Technology and Research Center, Alwar, India

Hari Shrawgi National Institute of Technology Raipur, Raipur, India

Kalpna Singh Centre for Interdisciplinary Research in Basic Sciences, Jamia Millia Islamia, New Delhi, India

Neha Singh Uttarakhand Technical University, Uttarakhand, India

Pushpendra Singh Electrical Engineering, JKLU, Jaipur, Rajasthan, India

Vaishali Singh Department of Computer Science, Jagannath University, Jaipur, India

Dilip Singh Sisodia National Institute of Technology Raipur, Raipur, India

Gunjan Soni Department of Mechanical Engineering, Malaviya National Institute of Technology Jaipur, Jaipur, Rajasthan, India

Foram Suthar Indrashil Institute of Science & Technology, Cadila Group, Ahmedabad, India

Akash Talwariya Electrical Engineering, JKLU, Jaipur, Rajasthan, India

D. Vasumathi Department of CSE, JNTU-H, Hyderabad, India

M. Venkatesh Saravanakumar PG and Research Department of Computer Science, Sudharsan College of Arts and Science, Pudukkottai, TN, India

Niti Vishwas National Institute of Technology, Agartala, India

Rohit Vyas Government College of Engineering & Technology, Bikaner, India

Real-Time Distributed Denial-of-Service (DDoS) Attack Detection Using Decision Trees for Server Performance Maintenance



Mrunmayee Khare and Rajvardhan Oak

Abstract With the incorporation of the Internet in our lives and its ever-increasing usage, it becomes all the more imperative to safeguard our systems and data against the malicious attacks. One of such malicious attacks is distributed denial-of-service (DDoS) attack, in which multiple nodes target a single target node to flood it. DDoS attacks in most of the cases directly target the server, thereby posing an important question on the security of the systems. DDoS attacks affect the entire network, thereby resulting in its downtime, unavailability of the services, and performance degradation. Due to this phenomenon, business losses are incurred, and hence, it becomes important to detect them before the damage is done. In this research work, a model is introduced which performs real-time detection of DDoS attacks. This model uses the decision tree classifier for detection of DDoS attacks. First, the features are extracted, and the information gain is calculated. Based on this, the decision tree is constructed which is then used to classify an instance as DDoS or not using a classifier algorithm. This model has demonstrated a success rate of 90.2% which is an improvement over the currently available algorithms. Furthermore, it also outperforms the existing algorithms in terms of sensitivity and specificity. The most significant feature of the model is that it operates in a live system in real-time conditions. This ensures protection against data losses and also helps in identifying the source of the attack. This system suggests a lightweight data mining approach to detect DDoS attacks using decision trees.

Keywords IDS · Cyber-crime · DoS · DDoS · Classification · Decision tree

M. Khare (✉) · R. Oak

Department of Computer Engineering, Pune Institute of Computer Technology, Pune, India
e-mail: khare.mrunmayee5696@gmail.com

R. Oak

e-mail: rajoak1995@gmail.com

© Springer Nature Singapore Pte Ltd. 2020
M. Pant et al. (eds.), *Performance Management of Integrated Systems and its Applications in Software Engineering*, Asset Analytics,
https://doi.org/10.1007/978-981-13-8253-6_1

1 Introduction

Distributed denial-of-service (DDoS) attack has been defined as an explicit attack by the attacker to exhaust the resources of the victim node [1]. Multiple nodes are deployed to launch the attack by sending a stream of packets toward the victim node, thus consuming the resources of the victim node and making them unavailable to the legitimate source nodes [1]. The attack can also be defined as an attempt by the attacker to flood the victim nodes with multiple requests and data, thereby making it unavailable to the legitimate nodes [2]. A denial-of-service attack (DoS) is defined when the attack is only done by one node in the network [3]. When multiple nodes attack the victim, the attack is defined as a DDoS [2, 4]. Thus, the victim node becomes unable to process so many packets (legitimate and attack) and denies service.

In a DDoS, an attacker node instructs several zombie nodes to flood the target with packets. DDoS attacks are mainly divided into two classes: bandwidth depletion [5] (where the goal of an attacker is to flood the victim node with a huge amount of traffic in order to prevent the legitimate traffic from reaching the victim node) and resource depletion [3, 5, 6] (where the goal of an attacker is to degenerate the critical resources of a victim node in order to prevent the legitimate user from using these resources) [3, 7].

DDoS severely affects the packet throughput for legitimate users, thereby increasing the packet loss, packet delay, and packet jitter [8].

In this paper, a novel system has been proposed and designed. This system aims at detecting the DDoS attacks at the destination node (victim node) in a live network, i.e., a real-time system. The algorithm runs continuously at the node and constructs decision trees for classifying the arriving packets. A count of the threshold is also maintained while doing so in order to prevent the attack.

This system uses the classifier algorithm to distinguish between the harmful packets and the legitimate packets. This ensures protection against data losses and also helps in identifying the source of the attack. This system suggests a lightweight data mining approach to detect DDoS attacks using decision trees.

2 Literature Survey

A survey of the existing mining techniques to detect a DDoS attack was conducted. Mining is carried out at the source or at the destination node. The source mining is deployed at the source of the attack in order to identify the attacker, whereas the destination mining is deployed at the victim node to detect the attack and trace the attacker. The authors in [8] and [3] have suggested a number of strategies for the same.

The intrusion-based mechanism in [9] extracts the required features from the database and applies two different techniques: Naïve Bayes and augmented Naïve

Bayes for intrusion detection. In [5], a traffic matrix is created using packet arrival time and source IP and variance is calculated to classify traffic as normal or DDoS. In [4], a decision tree which considers over 15 factors is deployed for attack detection and a traffic pattern matching technique for attack identification and its trace back. A Bayesian network-based technique proposed in [10] uses a statistical anomaly-based detection method based on KNN clustering [4]. Clustering is a set of techniques for finding patterns in data received from the packet capturing. This approach is used in [11]. The technique used in [12] uses transactional databases where each transaction has a temporal ID, a user ID, and an itemset. A sequence is an ordered list of itemsets, and the sequences are used to detect attacks.

In [7], a clustering-based technique has been proposed by the authors in which instances are clustered into malicious and non-malicious classes. Mohana Priya and Mercy Shalinie [13] suggests a mechanism to identify DDoS attacks in SDN using a restricted Boltzmann machine (RBM). Significant research has also been carried out in the domain of DDoS detection in cloud environments [14]. A stealthy, resilient IP filter-based system has been proposed in [15]. Due to extensive technology and modern tools available to attackers, application layer DDoS (App-DDoS) attacks have become a serious threat to Web server [16]. Hence, to deal with this, the authors in [17] have put forward two time series detection models: multifeatures information entropy predict model for flooding attacks and second-order Markov predict model for asymmetric attacks.

The strategies discussed in the literature [4, 5, 7, 9–13, 15] work on logs and hence do not perform live attack detection. Thus, the attack is identified once the damage is done. In addition, they consume a lot of space and memory and are complex to implement.

3 Proposed Model

3.1 Background Regarding Decision Trees

Decision trees classify data as per hierarchy and reaching the nodes to identify the nodes and machines used to flood the computer, i.e., the server. These results are then represented as a linear function that can effectively separate two types of classes when applied to the test data set.

Two algorithms are popularly used for the construction of decision trees: ID3 and C4.5.

Decision trees offer several advantages such as intuitive knowledge expression, high accuracy, and less memory consumption.

3.2 Proposed Model

The model consists of three phases: data capturing, preprocessing, and tree generation and classification. The various modules involved are as shown in Fig. 1.

The architecture of the proposed system is shown in Fig. 2. All packets which are classified as responsible for causing DDoS are then clustered to identify the attacking source node and thus to detect the DDoS and stop its further effects immediately.

Several choices are available for the classification phase such as decision trees, Naïve Bayes classifier, Bayesian belief networks, neural networks, or support vector machines (SVMs). In the proposed model, decision trees have been used. This is because they offer several advantages such as simple implementation, high accuracy, less memory requirement, and intuitive knowledge expression.

The model consists of three phases: data capturing, preprocessing, and tree generation and classification. The various modules involved are as follows.

- (1) *Data capturing phase:* This phase consists of obtaining data which consists of the logs captured by the network analysis tools like Wireshark and Ethereal. This data is then stored in the online database.
- (2) *Online database:* This is used to store the data in the form of its arrival time and along with all the fields from the packet capturer.

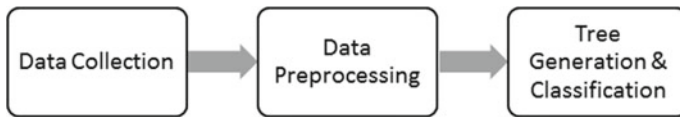


Fig. 1 Proposed model

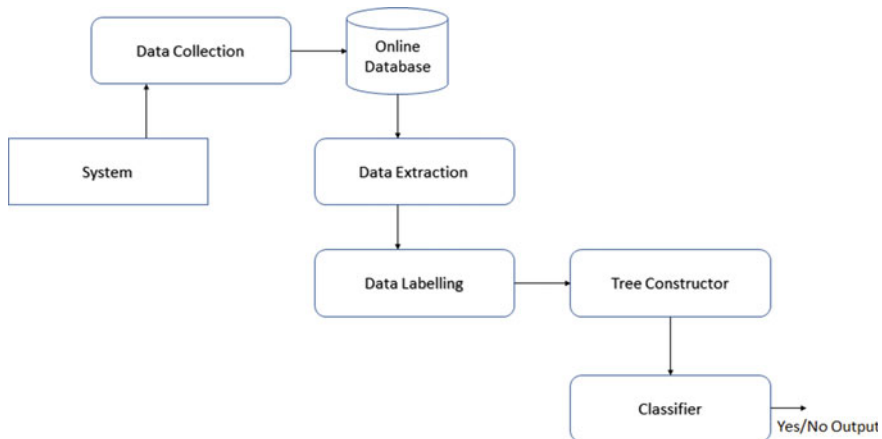


Fig. 2 Architecture of proposed system



- (3) *Extraction*: The data is then extracted, and only the required parameters like source address, arrival time, time to live, and destination address are maintained.
- (4) *Labeling*: The extracted data is then labeled which is used for tree construction by the classifier.
- (5) *Classifier*: The classifier builds a decision tree using C4.5 (yes–no) tree and at its leaf nodes gives the decision whether the packet is to be dropped or forwarded. It builds decision trees for each incoming packet. C4.5 offers significant advantages over ID3 such as higher speed, smaller tree size, and multiclass weighing.
- (6) *System*: It is the target node system from which data is being collected in real time.

3.3 Algorithm and Mathematics

This algorithm builds a decision tree using EVFDT [1] to differentiate between legitimate and malicious packets and then further uses the K -means clustering [11] to detect the source of the attack. This algorithm modifies the existing algorithm for better accuracy and efficiency. The following are the steps:

- (1) *Tree construction phase*: This phase is based upon the EVFDT [1] and has improvements in two parameters: accuracy and noise handling. This is obtained by calculating the bound values and information gain [5] and averaging the two for determining the node split instead of using the hefting values as used in EVFDT [1]. The tree building starts from the root node. If the size of the tree grows beyond the algorithm capacity, tree pruning [1] is implemented. Thus, the size of the tree is also handled. The bound value (best attribute) for splitting is calculated as follows:

$$GP(A, T) = \sum Si(A, T, v) + H(A, T) + \alpha$$

where

- $GP(A, T)$ is the bound value, for attribute A in a set of training samples T .
- $Si(A, T, v)$ is an individual entity in the set with value $= v$ for attribute A .
- $H(A, T)$ is the heuristic function determining of the information gain and splitting criterion as used in the EVFDT [1].
- α is the desired probability of choosing the correct attribute at a given node obtained by using the Bayes classifier [10].

- (2) *Clustering phase*: All the attack nodes as classified by the decision tree constructed were then clustered. A number of different approaches may be used [3, 7] such as hierarchical, K -means, and K -medoids. The proposed model has been implemented using K -means clustering [11] in order to identify the source of the attack.

The accuracy (1) of detecting the attack was calculated using the following

$$A = (TP + TN)/(n) \quad (1)$$

where

TP true positives, number of samples correctly classified as attack packets.

TN true negatives, number of samples correctly classified as legitimate packets.

FP false positive, number of samples incorrectly classified as legitimate packets.

FN false negatives, number of samples incorrectly classified as attack packets

Then, the parameters, sensitivity (2) and specificity (3), can be defined as follows:

$$\text{Sensitivity} = TP/(TP + FN) \quad (2)$$

$$\text{Specificity} = TN/(TN + FP) \quad (3)$$

This algorithm uses tree pruning as used in EVFDT [1] and thus is capable of building decision trees with reduced size, thereby increasing the capability of the algorithm to handle noise. The size denotes the depth of the decision tree, and EVFDT [1] produces the smallest decision trees among all the algorithms.

3.4 Analysis

Computation time is the total time in seconds for processing a full stream of data [1]. The time complexity of the proposed algorithm is

$$O(lav)$$

where l , a , and v denote length of the tree (pruned), total number of attributes, and values of attributes, respectively.

This algorithm takes more running time than its predecessors EVFDT [1] and VFDT- π [18]. The computation time for VFDT- π [18] as the size of π remains fixed.

Space complexity specifies the memory requirements of the algorithm. The space complexity of the following algorithm is:

$$O(nv)$$

where n , t , and v are number of decision nodes in the tree, total number of attributes, and values of attributes, respectively.

4 Results

The simulation of this model was performed using a network of 30 computer systems connected via LAN. One of the nodes was selected as the target. The remaining 29 machines were programmed to send continuous packets to the target IP. At the target, the algorithm was implemented to identify this DDoS. All machines were standard i3 processor-based and with Linux as operating system. The results obtained are shown in Table 1.

This algorithm can effectively build decision trees and detect the source of the attack in a real-time system. Due to the use of the prune mean tree strategy, it shows excellent accuracy even with the increase in the noise.

It was also observed that our system ranks better in terms of the parameters accuracy, sensitivity, and specificity. In the given Tables 2, 3 and 4, the values are presented by comparing it with average values for the various noise percentages (0, 5, 10, 15%). This system outranks the existing ones with respect to all parameters.

Table 1 Results

| Noise rate (%) | Sensitivity (%) | Specificity (%) | Accuracy (%) |
|----------------|-----------------|-----------------|--------------|
| 0 | 98.6 | 96.3 | 97 |
| 5 | 84 | 92.5 | 91.6 |
| 10 | 83.6 | 90.11 | 86.8 |
| 15 | 81.38 | 88.43 | 85.6 |

Table 2 Comparative study of accuracy values

| Algorithm | Accuracy (%) | Rank |
|--------------------|--------------|------|
| VFDT- π | 86.5 | 3 |
| EVFDT | 87.5 | 2 |
| Proposed algorithm | 90.2 | 1 |

Table 3 Comparative study of sensitivity values

| Algorithm | Sensitivity (%) | Rank |
|--------------------|-----------------|------|
| VFDT- π | 82.6 | 3 |
| EVFDT | 85.7 | 2 |
| Proposed algorithm | 87.3 | 1 |

Table 4 Comparative study of specificity values

| Algorithm | Accuracy (%) | Rank |
|--------------------|--------------|------|
| VFDT- π | 88.9 | 2 |
| EVFDT | 88.6 | 3 |
| Proposed algorithm | 89.8 | 1 |

5 Conclusion and Future Scope

Thus, after a detailed literature survey, a new architecture has been proposed along with an algorithm which is based upon EVFDT. On testing the implementation, an accuracy of 90.2% has been observed. In addition, there is less space overhead. However, there is a higher computation time. Real-time detection prevents the loss of data and hampering of resources.

The algorithm will be significantly improved if one can enhance it by reducing the computation time, while maintaining the same accuracy. In addition, future research may focus on maintaining accuracy at high scalability as well. The proposed system may be combined with other classifiers in a bagging or boosting multiclassifier system to achieve better results.

References

1. Latif, R., Abbas, H., Latif, S., & Masood, A. (2015). EVFDT: An enhanced very fast decision tree algorithm for detecting distributed denial of service attack in cloud-assisted wireless body area network. *Mobile Information Systems*, 2015, Article ID 260594.
2. Latif, R., Abbas, H., & Assar, S. (2014). Distributed denial of service (DDoS) attack in cloud-assisted wireless body area networks: A systematic literature review. *Journal of Medical Systems*, 38, Article 128.
3. Nikolskaya, K. Y., Ivanov, S. A., Golodov, V. A., Minbaleev, A. V., & Asyaev, G. D. (2017). Review of modern DDoS-attacks, methods and means of counteraction. In *2017 International Conference "Quality Management, Transport and Information Security, Information Technologies" (IT&QM&IS)*.
4. Wu, Y.-C., Tseng, H.-R., Yang, W., & Jan, R.-H. (2011). DDoS detection and traceback with decision tree and grey relational analysis. *International Journal of Ad Hoc and Ubiquitous Computing*, 7(2), 121–136.
5. Lee, S. M., Kim, D. S., Lee, J. H., & Park, J. S. (2012). Detection of DDoS attacks using optimized traffic matrix. *Computers & Mathematics with Applications*, 63(2), 501–510.
6. Hussein, S. M. (2016). Performance evaluation of intrusion detection system using anomaly and signature based algorithms to reduction false alarm rate and detect unknown attacks. In *2016 International Conference on Computational Science and Computational Intelligence (ICSCI)*.
7. Nikolskaya, K. Y., Ivanov, S. A., Golodov, V. A., & Sinkov, A. S. (2017). Development of a mathematical model of the control beginning of DDoS-attacks and malicious traffic. In *2017 International Conference "Quality Management, Transport and Information Security, Information Technologies" (IT&QM&IS)*.
8. Learning Methods for Buczak, A. L., & Guven, E. (2016). A survey of data mining and machine cyber security intrusion detection. *IEEE Communications Surveys & Tutorials*, 18(2), Second Quarter.
9. Najafi, R., & Afsharchi, M. (2012). Network intrusion detection using tree augmented Naive-Bayes. IEEE Iran Section.
10. Jensen, F. V. (2001). *Bayesian networks and decision graphs*. New York, NY, USA: Springer.
11. Jain, K., & Dubes, R. C. (1988). *Algorithms for clustering data*. Englewood Cliffs, NJ, USA: Prentice-Hall.
12. Agrawal, R., & Srikant, R. (2008). Mining sequential patterns. In *Proceedings of IEEE 11th International Conference on Data Engineering* (pp. 3–14).

13. Mohana Priya, P., & Mercy Shalinie, S. (2017). Restricted Boltzmann Machine based detection system for DDoS attack in Software Defined Networks. In *2017 Fourth International Conference on Signal Processing, Communication and Networking (ICSCN)*.
14. Manoja, I., Sk, N. S., & Rani, D. R. (2017). Prevention of DDoS attacks in cloud environment. In *2017 International Conference on Big Data Analytics and Computational Intelligence (ICBDAC)*.
15. Aline Sophia, G., Gandhi, M. (2017). Stealthy DDoS detecting mechanism for cloud resilience system. In *2017 International Conference on Information Communication and Embedded Systems (ICICES)*.
16. Jiang, M., Wang, C., Luo, X., Miu, M. T., & Chen, T. (2017). Characterizing the impacts of application layer DDoS attacks. In *2017 IEEE International Conference on Web Services (ICWS)*.
17. Wang, Y., Liu, L., Si, C., & Sun, B. (2017). A novel approach for countering application layer DDoS attacks. In *2017 IEEE 2nd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*.
18. Hulten, G., Spencer, L., & Domingos, P. (2001). Mining time-changing data streams. In *Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'01)* (pp. 97–106), San Francisco, CA, USA, August 2001.
19. Zargar, S. T., Joshi, J., & Tipper, D. (2013). A survey of defense mechanisms against distributed denial of service (DDoS) flooding attacks. *IEEE Communications Surveys & Tutorials*, 15(4), 2046–2069.
20. Thwe, T., & Thandar, P. (2014). Statistical anomaly detection of DDoS attacks using K-nearest neighbour. *International Journal of Computer & Communication Engineering Research*, 2(1), 315–319.
21. Narasimha Mallikarjunan, K., Muthupriya, K., & Mercy Shalinie, S. (2016). A survey of distributed denial of service attack. In *2016 10th International Conference on Intelligent Systems and Control (ISCO)*.

Cloud Computing: Vulnerability and Threat Indications



Vaishali Singh and S. K. Pandey

Abstract An innovative approach of providing and delivering resources is the new emerging technology “cloud computing.” The time demands a reduction in expenditure as an end result of financial restrictions, where cloud computing has established productive position and is seeing immense large-scale investment. Still, despite the significance, users and clients are nervous at the prospects and security issues of cloud technology. Threat and vulnerability factors are majorly, one of the scrutinizing issues in cloud, if it is not properly secured due to which, a direct control loss over the system is creating nevertheless accountable threat. This paper explains, based on present scenarios, how the vulnerabilities analysis will evaluate current emerging threats of cloud technology. The paper focuses to provide indication of the recent threats and vulnerabilities in cloud. The study will help the providers and users to make knowledgeable decisions about threat mitigation and threat countermeasures within a cloud strategy.

Keywords Vulnerabilities · Threats · Security · Cloud computing · Cloud security

1 Introduction

Cloud is characteristically represented as computing services that depend on allocation of resources in spite of having confined servers or individual devices to hold applications. Cloud can be well thought-out as a symbol of the Web sphere (*Internet*) [1]. The reason behind using cloud as a representation of Internet is that cloud provides effective computing by integrating repositioning of data, processing, and

V. Singh (✉)

Department of Computer Science, Jagannath University, Jaipur, India
e-mail: vaishalisingh@stxaviersjaipur.org; vaishali.siingh@gmail.com

S. K. Pandey

Department of Electronics & Information Technology, Ministry of Communications & IT,
Government of India, New Delhi, India
e-mail: santo.panday@yahoo.co.in

confidential information handling procedures [1]. Most common case in point of cloud technology is the use of e-mail services provided by many social networking sites; in this setup, the cloud manages the e-mail management software and Web servers which is completely managed by the service provider on virtual machines giving rise to so-called virtualization [2].

The massive use of virtualization in executing cloud computing infrastructure brings security alarms for several services provided by cloud [2]. Various but very significant security challenges are measured while using and implementing virtualization for cloud computing [3]. Some of the security concerns are like undetected network threats and attacks, allocating assets, and de-allocating resources, and virtual machine (VM) hypervisor has been seen. Like cloud virtualization, there are different challenges and issues associated with cloud computing [3].

Cloud issues and challenges can be clustered into various magnitudes like security comprising confidentiality, compliance challenges, and legal issues [4, 5]. Process and methods used for making security measure available to any users are now hidden behind levels of abstraction, creating various cloud security issues [4, 5]. These security issues generally consist of identifying vulnerabilities, threats, and attacks with their countermeasures so that security can be provided at each layer of cloud technology in the outline of ontology [6, 7]. Regardless of providing various countermeasures for these cloud security issues, cloud services are required to be more protected and well-built to fulfill the daily needs of the clients [5]. So, there is a requirement of analyzing the vulnerabilities factor of these securities issues and providing countermeasures for them [8].

Vulnerability is any potential circumstances or event that could destructively affect a project's ability. Vulnerability is a revelation to failure or injury or issue, element, or thing, which have uncertain danger [9]. Information technology (IT) in many businesses management plays a significant role. In the present world where information technology (IT) plays a very considerable work in the development of any business organization, identification and study of vulnerability factors are very essential. Once the vulnerability factors are identified, then only proper countermeasures can be set to avoid and minimize them.

Further than this preface on the background facts, the remnant of the research work is structured as follows. Section 2 highlights the "Vulnerabilities in Cloud Computing." Later on, "Ten top emerging Cloud Threats" is given in the Sect. 3. To end with, conclusion and future work are outlined in Sect. 4.

2 Vulnerabilities in Cloud Computing

Cloud industry is at its boom; it is expected that around 2014, cloud will turn out to be a \$150 billion trade [10]. The reason behind this success is that whether the user is at desktop or laptops or mobile device, etc., they will have the rights to access their data anytime, anywhere with an high-speed uninterrupted Internet (www) connection [10].

Cloud computing also offers good services to business world as extensible data storage for files and improved teamwork among members irrespective of geographical locations, by saving their time and money and by eliminating the requirement of building an expensive data center and an IT expert to manage the business affairs [10].

But all these facilities of cloud become vague in the presence of biggest issue, i.e., the security. Even though most trustworthy cloud service providers (CSPs) have unmatched security to protect user's data, then also some of the experts are of the view that no secure cloud system till date exists.

The following are the main security vulnerability issues present in cloud system which all users should understand when backing up or storing their assets in the cloud.

- i. **Third-party watchman:** Cloud is totally different from data centers, where data are controlled and supervised by self-controlling IT department. In cloud, it is the cloud service providers (CSPs) who enjoy the overall control on the users' data [11, 12]. Provider does the whole fixation from performing all system updates and maintenance to managing cloud security. Broadly speaking, it is the third party whom the user completely trusts to keep their data secure [11, 12].
- ii. **Cyber-attacks:** Storing data within the cloud creates chances of cyber-attacks vulnerabilities. This vulnerability becomes more prominent in cloud where many users store their diverse data on the same cloud system. The terrifying thing is the vulnerability to distributed denial-of-service attacks [10]. It is rightly surveyed by many researchers that cloud has a solitary point of failure which means if as lightest of a thing gets wrong, it lays an impact on a very vast group of people. Corrupting bulk of data is an easy task than to deal with single data [10, 13]. While cloud service providers (CSPs) have many security measures, still the hacker is more technology-experienced. It is generally assumed that cyber-criminals do not attack the entire cloud, but they work on small domains like hacking of user's account [13].
- iii. **Insider threats:** Breaches result in loss of information. Threats are not only caused by outsider, but insider also plays a very imperative role. The insider can be any member of cloud service providers who can create dangers for any user. Once there occurs malicious access toward the customer's data, confidential information and intellectual properties are up for grabs [14]. In cloud, threat from insider is very crucial as it worsens the security factor by giving the cloud management platform their administrative rights, either by an employee or by an attacker who poses as an employee, grants access to copy and steal any virtual machine, undetected, or potentially infects and destroys the entire cloud environment in couple of minutes [14].
- iv. **Government intrusion:** According to recent survey, not only the third party but also the government keeps a keen eye on the third-party data kept in the cloud system. The government can any time view the user data if the data transmitted outside geographical region are susceptible [15]. So, dissatis-

fied customer's data are always at vulnerability from competitors, workforce member breaching security of cloud, and government intrusion.

- v. **Legal liability:** Vulnerabilities within the cloud are partially related to security breaches. But it also includes lawsuits filed by or against the user. According to survey, the current vulnerability in cloud via services is authorized liability and business continuity [16]. The benefits of cloud are to be weighed against the extent of security measures irrespective of its ease of access, collaboration, and rapidity, and it is rightly said that information security has constantly been concluding an equivalent set of scales between ease of access and sharing verses entirely protected down security [16].
- vi. **Lack of standardization:** Safety of cloud plays a very prominent role in today's era. The need of cloud standardization is required as no specific vulnerability assessment measure for cloud exists. Further, it becomes problematic for users to conclude accurately how safe and secure the cloud really is [17]. The answer to the question "How much safe the cloud is?" has many aspects, and its answer depends on cloud service provider, the category of industry a company is in, and the associated regulations regarding the data it is considering cloud storing. The safety in cloud varies from dissimilar service providers [17].
- vii. **Data Availability and Business Continuity:** Cloud service providers try to build a good security interface for their customers, but despite all the efforts, the consumers are still facing a major vulnerability as they never know when the data will not be accessed by them due to loss of Internet connectivity [18]. Some providers may have feasible customer support, but not all. Customer support provided by cloud service providers plays a very important role in organizations dealing with critical issues where slight mismanagement is not appreciable [18].
- viii. **Availability:** The most common vulnerability in cloud is faced due to unaware redundancy and fault tolerances is not under the user's control. The services provided within cloud have no fragile fault acceptance and absolute availability, yet the users are not able to access their data for longer period of time with service interruptions not only this the third party and the hackers have a constant view on the user's confidential and private data [19]. The cloud providers generally perpetrate that they do remarkable, tripartite-protected information backups, but then also users have lost data permanently. As new innovation is made in information technology, so new methods and chances of vulnerabilities have cropped in the IT sector. So, it is necessary that the customers should have backup of the data which is shared with the cloud or at slightest it can claim on legalese that has the correct quantity of damages built-in if those records are gone forever [19].
- ix. **Authentication, Access Control, and Authorization:** Software usage is incomplete without authentication, authorization, and access control (AAA) which is a very crucial mechanism [20]. The service providers should setup types of rights of users, if in any condition violated than the user's account

can be deactivated and all privileged provided to the user, all this should be well stated and mentioned in Service level Agreement (SLA) [20].

- x. **Data Security and Privacy:** Privacy and security both are complementary to each other. Both play a very important and effective role in cloud computing, which is a matter of huge concern. The user should understand the methods and measures undertaken by the cloud service providers (CSPs) to safeguard and outsource their confidential data [21]. The service-level agreement (SLA) should have the detailed description of all security and mitigation measure and government policies. The cloud service provider (SLA) should be well versed with some of the data security and data privacy rules and regulations that apply to the user's data entity, such as Payment Card Industry, Data Security Standard, and the Federal Information Security Management Act of 2002.
- xi. **Location of data:** For the users, it is essential to identify the location of storage of their confidential data. This makes the user aware to know where the data are stored and what security/privacy rules and policy are applicable to it, as policies and rules differ from region to region [22].
- xii. **Disaster recovery:** When a cloud service provider (CSPs) hosts users' confidential and crucial data, it becomes immensely essential to have disaster recovery capabilities and disaster recovery strategy [23].
- xiii. **Shared access:** The main focus of cloud is sharing the same storage area and computing assets and resources like CPU, memory, namespace, and physical building and many more with multiples users. This creates additional vulnerabilities while sharing recourses within unknown users [21]. The greatest threat to cloud is other customers or malicious users can see all other data and can pilfering the identity of another client. This gives rise to several new classes of vulnerabilities, and users having malicious intention can peak into tenants' memory and IP address space [21].
- xiv. **Virtual exploits:** Cloud is a substitute of virtualization, and a service provider is a user of virtualization. There are numerous unique threats which goal the virtual server hosts and the users. All together virtually exploited vulnerabilities are mostly unknown and uncalculated in most popular vulnerability models the reason behind this is because of client's usually has no awareness and understanding about virtualized goods or management tools which the vendor/provider is running [24].
- xv. **Ownership:** The user of cloud faces surprise vulnerability as who owns the data. Generally, the user is unaware of the ownership of their personal and confidential data. Ownership of cloud data is vulnerable as the service providers can even misuse the user's data for personal gains.
- xvi. **Short-term negotiation on contracts:** Another vulnerability the cloud consists of is that the customer is entirely at the mercy of the terms and conditions offered by cloud providers. The service providers set the contracts and the conditions of agreement around key issues like service-level availability, and the user has left with the only choice to accept it. Some providers may discuss the

customer choice, but majority imposes service-level agreement (SLA) made by them only.

- xvii. **Difficulty in creating hybrid systems:** This applies to organizations holding sensitive and penetrating information like government offices and financial institutions. They usually have their own IT departments and will not take their data to cloud regardless of the benefits of efficiency and performances of cloud service providers (CSPs). These cloud service providers (CSPs) generally follow no specific standards, and the chance of reliability reduces for organization dealing with sensitive data to put their data on cloud.
- xviii. **Centralization:** Centralized data can undoubtedly add another vulnerability to cloud computing. It means that if the cloud service provider server system goes down, all the clients will be affected.

3 Ten Top Emerging Cloud Threats

Security is being a continuous issue for cloud computing technologies. The cloud implementation in business, lack of privacy, and security is the foremost problem. This has directed a way to enhance the challenges in the cloud. This work identifies the rising threats in cloud. These threats are given as follows and also shown in Fig. 1.

The following are the top 10 recent threats:

- i. **Repudiation:** Due to the presence of integrity vulnerabilities, repudiation has been aroused as a key problem in the current usage of cloud storage platforms. Repudiation can be very well defined as tampering or modification of facts so that the former party is unaware of the original information. This problem works as a good and effective tool for the persons who are taking advantages of someone information like blackmailers and intruders. Not only the outsiders but also the insider in any organization also helps in leaking the valuable information. Due to this, integrity and repudiation is one of the foremost problems faced by many organizations resulting in mismanagement on various storage platforms on cloud. For several years, nearly all the organizations have faced problems due to repudiation threat, and without adequate auditing, but still the issues are the same? Non-repudiation protocol is also been used but unable to resolve the difficulty at a greater extent.
- ii. **Replay Attack:** Another major problem faced in cloud platforms is replay attack. In this problem, both the sender and receiver are unaware of interference of information by the third party. The solution recommended to this problem was to initiate the concept of time stamp. Time stamp requires synchronization of timing at both the sender and receiver ends which is not possible in distributed cloud environment. So many protocols use randomly generated nonce values to avoid replay attacks. These nonce values are uniquely generated for every session, and the receiver will be able to recognize a replay of the earlier send message holding an old nonce value.



Fig. 1 Cloud threats evaluated through vulnerability analysis

- iii. **Identity Forgery/Spoofing Attack:** When the authorized identity tokens or statements are manipulated or altered deliberately by malicious persons, it is known as identity forgery/spoofing attack. This leads to no disclosure of original facts and figures leading the investigation to wrong directions. When working with cloud-based IDMS, such types of forgery can be identified by indicating strict (two-factor) authentication mechanisms. These types of forgery lead to fraud and identity theft and may require expert knowledge and exceptional skills to solve it.
- iv. **Luring Attack:** Same as elevation of privilege attack, a new attack has emerged known as luring attack in which an invader might interest a superior privileged guest to get an action on his or her behalf. In the infrastructure as a service model, if the attacker is capable to lure the method to stop virtualization, it will be a big operational vulnerability. In this attack, the attacker lures a module which is highly privileged to do something on the behalf of the client. Further, the attacker convinces the client to execute the attacker's code in a straightforward method in additional security context. The result is not an authorization malfunction but relatively a breakdown of the system which is not appropriately notified to the user. The cloud identity management system is more vulnerable to luring attack as it does not provide logging and reporting

mechanism. Further specifically, the challenger lures highly privileged client to execute several illicit actions on their respect.

- v. **Elevation of Privilege:** The attacker takes the help of programming errors or designing faults in privilege escalation attacks to misguide the data and application accessible on the networks. It is generally needed in that circumstance where system does not allow any unauthorized user to enter the system at the primary level. When the attacker tries to use the kernel to enter the system, he generally gets the permission of administrator and enjoys a higher privilege. This kind of attack is recognized as vertical privilege escalation. In horizontal privilege escalation the attacker enjoys the same rights similar to the level of identity of the authorized user like if anyone gains the access to a authorized person's bank account, he will enjoy the rights of that person using horizontal privilege escalation attack. In this, the attacker gets limited attack access rights. This category of attack is generally found in cloud models as they provide security to the user with limited privileges in which the attacker may also enjoy. The users of IDMS may enjoy restricted set of privileges in privilege escalation attack which may even cause harms to stored information of cloud.
- vi. **Network Threats:** A variety of threats like man-in-the-middle attack, network sniffing, SQL injection attack, cross-site scripting, and many more are being detected as network threats. Man in the middle is a type of attack in which all of the communications can be tracked by the third party if secure socket layer is not appropriately being installed, whereas in network sniffing, the plain text is tracked on the network. Cross-site scripting is a form of attack where the URL is redirected on the attackers to and confidential information is traced, whereas in SQL injection attack, the attacker modifies the query by adding additional information.
- vii. **Canonicalization Attacks:** In canonical attacks, the inputs have the same name, so the code is more susceptible to attacks. This kind of problem occurs if the security decisions are basically focused on the name of the resources that are sent to the programs as inputs. Generally, paths, files, and URLs are susceptible to canonicalization because each may have many ways to represent the same name.
- viii. **Dictionary Attack:** This type of attack is very commonly used by the spammers. They get all the details of the user like name and password and try to access the system. These types of attacks are generally avoided and unsuccessful against the system which uses multiple word phrases.
- ix. **Misuse of Infrastructure:** Infrastructure also plays an imperative responsibility in cloud. If the cloud data are not well facilitated with authoritarian requirement, then also security breaches take place. Authoritarian requirement means the legal aspects of any data like who is the owner of the data, what are the legal formalities associated with that data, and what type security breaches related to that data if something went wrong to will be held responsible for it.
- x. **Credential Theft:** This kind of theft is mostly related to the hacking of the verification part of the confidential information. Passwords are considered to be the most important element of credential theft, so if password is leaked or

revealed to the attacker, then the system can be easily tampered for which this password was used. This may keep the data on cloud at a very high vulnerability not only for the provider but also for the client.

4 Conclusion and Future Work

Security is a vigorous part of study under the concern of cloud computing. Less work has been accomplished. Still the research is going on cloud on various issues and challenges. The study highlights a variety of cloud vulnerabilities and threats that at present affect the cloud computing environment. However, there may be a number of more security threats and vulnerabilities. Research study is currently undertaken on the diverse known threats and vulnerabilities faced across by the cloud and probable solutions for the same.

Despite of these research findings, there is a vital necessitates to work on the security domain to come up with the original thoughts associated with the mitigating measures. In this study, we have provided an outline of the ten main threats and vulnerabilities under a variety of areas, which may serve an initial step to discover the countermeasures of the listed ones. The study can provide a significant support to the related research areas, where further work is necessary particularly for entry-level researchers.

Future study may be to outline these recognized threats with reference to their various related parameters like security requirements and countermeasures. In addition, these threats can be embedded in the cloud security ontology to represent the facts and findings in more specific scientific way. Moreover, the appropriate mitigating technique may also be developed next to each cloud threat to present adequate security to both the user and service provider. The research work will lead cloud to a better confidence among the stack holders.

References

1. Pandey, S. K. (2013). Cloud computing: A new era of information technology management. *The Chartered Accountant Student (Students' Journal)*, SJ 4(1), 10–12.
2. Virtualization and Cloud Computing Steps in the Evolution from Virtualization to Private Cloud Infrastructure as a Service, Intel IT centre, August 2013. <http://www.intel.in/content/dam/www/public/us/en/documents/guides/cloud-computing-virtualization-building-private-iaas-guide.pdf>.
3. Singh, V., & Pandey, S. K. (2013, August). Research in cloud security: Problems and prospects. *International Journal of Computer Science Engineering and Information Technology Research (IJCSEITR)*, 3(3), 305–314.
4. Singh, V., & Pandey, S. K. (2013, July). Revisiting cloud security issues and challenges. *International Journal of Advanced Research in Computer Science and Software Engineering*, 3(7), 1–10.

5. Singh, V., & Pandey, S. K. (2013, September). Cloud security related threats. *International Journal of Scientific & Engineering Research*, 4(9), 2571.
6. Singh, V., & Pandey, S. K. (2014, November). Revisiting security ontologies. *International Journal of Computer Science Issues*, 11(6, No. 1), 150–159.
7. Singh, V., & Pandey, S. K. (2014). A comparative study of cloud security ontologies. In *2014 IEEE 3rd International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)* (pp. 797–803).
8. Vulnerability analysis framework for a cloud specific environment, Atos Company, White Paper. <https://atos.net/content/dam/global/we-do/atos-cloud-vulnerability-analysis-white-paper.pdf>.
9. Wooley, P., et al. (2011). *Identifying cloud computing security vulnerabilities* (p. 74). Capstone Report. University of Oregon Applied Information Management program.
10. CRN Staff, Cloud computing services market to near \$150 billion in 2014. <http://www.crn.com/news/managed-services/225700984/cloud-computing-services-market-to-near-150-billion-in-2014.htm>.
11. Butler, B. (2013, September 30). *Cloud Security Alliance's new guidelines focus on mobile, data management*. <http://www.networkworld.com/article/2170364/cloud-computing/cloud-security-alliance-s-new-guidelines-focus-on-mobile-data-management.html>.
12. Beckham, J. (2011, May 3). *The top 5 security risk of cloud computing*. <http://blogs.cisco.com/smallbusiness/the-top-5-security-vulnerabilities-of-cloud-computing>.
13. Shue, C. A., & Lagesse, B., *Embracing the cloud for better cyber security*. <http://faculty.washington.edu/lagesse/publications/cloudsecurity.pdf>.
14. Kandias, M., Virvilis, N., Gritzalis, D., *The insider threat in cloud computing*. <http://www.cis.aueb.gr/Publications/CRITISCloud%20Insider.pdf>.
15. Information Security Breaches Survey 2013 | technical report. <https://www.pwc.co.uk/assets/pdf/cyber-security-2013-technical-report.pdf>.
16. Weber, R. H., & Staiger D. N. (2014). Cloud computing: A cluster of complex liability issues. *Web JCLI*, 20(1).
17. Ortiz, S., Jr. (2011, September). *The problem with cloud-computing standardization*. <http://www.infoq.com/articles/problem-with-cloud-computing-standardization>.
18. Kleyman, B. (2014, 20 October) Combining cloud with disaster recovery and business continuity. <http://www.datacenterknowledge.com/archives/2014/10/20/combining-cloud-disaster-recovery-business-continuity/>.
19. Silva, P., *Availability and the cloud*, F5 White Paper, <http://www.f5.com/pdf/white-papers/availability-cloud-wp.pdf>.
20. Ranjith, D., & Srinivasan, J. (2013, May). Identity security using authentication and authorization in cloud Computing. *International Journal of Computer & Organization Trends*, 3(4).
21. Brodtkin, J., Gartner: Seven cloud-computing security risk. Available at: www.infoworld.com/d/security-central/gartner-seven-cloud-computing-security-vulnerabilities-853. See more at: <http://www.cepis.org/index.jsp?p=641&n=825&a=4758#sthash.TZhfG2f.dpuf>.
22. Kumar, P., & Singh, H. (2013). Data location in cloud computing. *International Journal for Science and Emerging Technologies with Latest Trends*, 5(1), 24–27.
23. Wood, T., Cecchet, E., Ramakrishnan, K. K., Shenoy, P., van der Merwe, J., & Venkataramani, A., *Disaster recovery as a cloud service: Economic benefits & deployment challenges*. <http://lass.cs.umass.edu/papers/pdf/HC10-dr-cloud.pdf>.
24. Vic (J.R.) Winkler, Cloud computing: Virtual cloud security concerns. <http://technet.microsoft.com/en-us/magazine/hh641415.aspx>.

Proposed Algorithm for Creation of Misuse Case Modeling Tree During Security Requirements Elicitation Phase to Quantify Security



Ajeet Singh Poonia, C. Banerjee, Arpita Banerjee and S. K. Sharma

Abstract Gathering secure measurement is the first step toward the developed of comprehensive secured software. Security is an intangible measure also considered as a non-functional attribute which needs to be quantified in some manner using tools and techniques during the preliminary phases, i.e., the requirements engineering stage of software development process. Studies carried out so far have shown and suggested that among the available techniques of security measurement, misuse use case modeling which is a form of unified modeling approach is very easy to implement during the requirements engineering (RE) phase of SDLC. This research work proposes an algorithm for creation of misuse case modeling tree during the security requirements elicitation phase. This algorithm can be customized according to the specific software application and is supported and synchronized with industry accepted standards like Common Vulnerability Scoring System (CVSS) and Common Vulnerability Enumeration (CVE). The work proposed is an extension of Misuse Case Oriented Quality Requirements (MCOQR) Framework and metrics and includes software application-specific database. The proposed work also showcases the areas where future work can be carried out to further fortify the entire system during the software development process, thereby contributing to the enhancement of various measures of performance.

A. S. Poonia

Government College of Engineering and Technology, Bikaner, India

e-mail: pooniaji@gmail.com

C. Banerjee (✉)

Amity University Rajasthan, Jaipur, India

e-mail: chitreshh@yahoo.com

A. Banerjee

St. Xavier's College, Jaipur, India

e-mail: arpitaa.banerji@gmail.com

S. K. Sharma

Modern Institute of Technology and Research Center, Alwar, India

e-mail: sharmasatyendra_03@rediffmail.com

Keywords Vulnerability · Misuse cases · CVSS · CVE · Security requirements elicitation phase

1 Introduction

One of the fundamental principles of engineering is measurement as it gives some tangible value for analysis purpose [1]. Measurement can be thought of as a raw data produced by counting certain factors which could be specific or discrete by nature [2]. These measurements themselves might not reveal any specific information unless it is correctly analyzed [3]. These analyzed measurements are called metrics and may exhibit quantitative or qualitative measures. Researchers suggested that if an activity cannot be measured, then it cannot be managed and remains ineffective if not analyzed in a proper manner [4].

Software also contains measurements, and certain metrics could be formulated around them which could give an insight into software's security and helps to uncover and explore problematic areas so that proper remedial measure could be developed and implemented. Hence, metrics are indicators and estimators of specific factors or certain aspects of a system which could suggest further improvement.

Similarly, security metrics could be devised to estimate the security level of software, so that, proper countermeasures could be suggested for the improvement of its security level. Hence, software security metrics provide a standard means for measurement of software security to define the level of security, the performance of the security aspects, and various indicators of security, i.e., its strength [5, 6].

In software development process, there are a number of metrics which are used for various purposes, viz. size-oriented metrics which works on the concept of line of code (LOC) and provides various indicators and estimators like defect per KLOC, cost per KLOC, manpower hours per KLOC, function-oriented metrics [7], which works according to function points as a normalization value, quality metrics which works on defect removal efficiency factor, process metrics which is used to gather various indicators and estimators of a process, project metrics which is used to gather various indicators and estimators of a project, and many more but none of them can be used or extended for the quantification and implementation of security in software [8, 9].

This research work proposes an algorithm for creation of misuse case modeling tree during the security requirements elicitation phase. This algorithm can be customized according to the specific software application and is supported and synchronized with industry accepted standards like Common Vulnerability Scoring System (CVSS) and Common Vulnerability Enumeration (CVE). Apart from Sect. 1 which mainly focuses on introduction, the rest of the research work contains the following: Sect. 2 presents the proposed algorithm for creation of misuse case modeling tree

Table 1 MCOQR framework and derived metrics scoring and predict degree of security of the proposed application (based on CVSS scoring and ranking standards)

| Scoring obtained from misuse case modeling tree | Predicted degree of security of the proposed application (based on CVSS scoring and ranking standards) |
|---|--|
| 0.0 | None |
| 0.1–3.9 | Low |
| 4.0–6.9 | Medium |
| 7.0–8.9 | High |
| 9.0–10.0 | Critical |

with count, scoring, and ranking, Sect. 3 discusses the implementation mechanism of the proposed algorithm, Sect. 4 presents the results and discussion, whereas in Sect. 5, the conclusion and future research work are given.

2 Proposed Algorithm

This section shows the proposed algorithm for the creation of misuse case modeling tree with count, scoring, and ranking (Table 1).

```

Step 1  Open  VULNERABILITY_DATABASE  alias  as  VD,
        CVSS_METRICS_DATABASE  alias  as  CVSSMD,
        MISUSE_CASE_DATABASE  alias  as  MCD,
        APPLICATION_SPECIFIC_MISUSE_CASE_DATABASE  alias  as
        ASMCD
Step 2  Declare variable 'n' and 'm' as integer and initialize to value '0'
Step 3  Declare variable 'flag1', 'flag2', 'V_ID', 'MC_ID' as string and
        initialize to value SPACES
        Do While .NOT. VD→EOF()
            Read VD→CURRENT_RECORD
            m = m + 1
        GoTo VD→NEXT_RECORD
        Loop
        End Do
Step 4  Declare dynamic array IIM_COUNT[m][3], IIM_SCORING[m][3],
        IIUM_COUNT[m][2], IIUM_SCORING[m][2], INIM_COUNT[m][3],
        INIM_SCORING[m][3], INIUM_COUNT[m][2],
        INIUM_SCORING[m][2], EIM_COUNT[m][3],
        EIM_SCORING[m][3], EIUM_COUNT[m][2],
        EIUM_SCORING[m][2], ENIM_COUNT[m][3],
        ENIM_SCORING[m][3], ENIUM_COUNT[m][2],
        ENIUM_SCORING[m][2] and initialize all values to '0'
Step 5  GoTo VD→BOF()
        Do While .NOT. VD→EOF()
Step 6      Read VD→CURRENT_RECORD→Vulnerability_Type_ID and Store in 'V_ID'
Step 7      Do While .NOT. ASMCD→EOF()
Step 8          Read ASMCD→CURRENT_RECORD→Vulnerability_Type_ID and Match with 'V_ID'

```

```

Step 9      If Match Found Then Read ASMC $\rightarrow$ CURRENT_RECORD $\rightarrow$ Misuse_Case_ID and Store in
            'MC_ID'
Step 10     GoTo MCD $\rightarrow$ BOF()
            Do While .NOT. MCD $\rightarrow$ EOF()
Step 11     Read MCD $\rightarrow$ CURRENT_RECORD $\rightarrow$ Misuse_Case_ID and Match with 'MC_ID'
Step 12     If Match Found Then Read MCD $\rightarrow$ CURRENT_RECORD
Step 13     If MCD $\rightarrow$ CURRENT_RECORD $\rightarrow$ EMetrics_Scoring $\leq$ 3.9 Then flag1='Non_Intrusive'
            Else flag1='Intrusive'
            EndIf
Step 14     GoTo CVSSMD $\rightarrow$ BOF()
            Do While .NOT. CVSSMD $\rightarrow$ EOF()
Step 15     Read CVSSMD $\rightarrow$ CURRENT_RECORD
            If CVSSMD $\rightarrow$ CURRENT_RECORD $\rightarrow$ Metrics_ID =
            MCD $\rightarrow$ CURRENT_RECORD $\rightarrow$ BMetrics_ID1 Then

                If CVSSMD $\rightarrow$ CURRENT_RECORD $\rightarrow$ Metrics_ID = 'BM001' or
                CVSSMD $\rightarrow$ CURRENT_RECORD $\rightarrow$ Metrics_ID = 'BM002' Then flag2='Internal'
                Else If CVSSMD $\rightarrow$ CURRENT_RECORD $\rightarrow$ Metrics_ID = 'BM003' Then
                flag2='External'
                EndIf
            EndIf
            Exit Loop
            Else
            GoTo CVSSMD $\rightarrow$ NEXT_RECORD
            Loop
            EndIf
Step 16     End Do
Step 17     GoTo CVSSMD $\rightarrow$ BOF()
            Do While .NOT. CVSSMD $\rightarrow$ EOF()
Step 18     If CVSSMD $\rightarrow$ CURRENT_RECORD $\rightarrow$ Metrics_ID =
            MCD $\rightarrow$ CURRENT_RECORD $\rightarrow$ TMetrics_ID2 Then
                If CVSSMD $\rightarrow$ CURRENT_RECORD $\rightarrow$ Metrics_ID = 'TM006' Then
                If flag1='Intrusive' and flag2='Internal' Then
                IIM_Count[n][0]=IIM_Count[n][0]+1
                IIM_Scoring[n][0]=IIM_Scoring[n][0]+MCD $\rightarrow$ CURRENT_RECORD $\rightarrow$ 
                EMetrics_Scoring
                Else If flag1='Non-Intrusive' and flag2='Internal' Then
                INIM_Count[n][0]=INIM_Count[n][0]+1
                INIM_Scoring[n][0]=INIM_Scoring[n][0]+MCD $\rightarrow$ CURRENT_RECORD $\rightarrow$ 
                EMetrics_Scoring
                Else If flag1='Intrusive' and flag2='External' Then
                EIM_Count[n][0]=EIM_Count[n][0]+1
                EIM_Scoring[n][0]=EIM_Scoring[n][0]+MCD $\rightarrow$ CURRENT_RECORD $\rightarrow$ 
                EMetrics_Scoring
                Else If flag1='Non-Intrusive' and flag2='External' Then
                ENIM_Count[n][0]=ENIM_Count[n][0]+1
                ENIM_Scoring[n][0]=ENIM_Scoring[n][0]+MCD $\rightarrow$ CURRENT_RECORD $\rightarrow$ 
                EMetrics_Scoring
                End If
            End If
            End If
            End If

```

```

Else If CVSSMD→CURRENT_RECORD→Metrics_ID = 'TM007' Then
  If flag1='Intrusive' and flag2='Internal' Then
    IIM_Count[n][0]=IIM_Count[n][1]+1
    IIM_Scoring[n][0]=IIM_Scoring[1][0]+MCD→CURRENT_RECORD→
    EMetrics-Scoring
    Else If flag1='Non-Intrusive' and flag2='Internal' Then
      INIM_Count[n][1]=INIM_Count[n][1]+1
      INIM_Scoring[n][1]=INIM_Scoring[n][1]+MCD→CURRENT_RECORD→
      EMetrics-Scoring
    Else If flag1='Intrusive' and flag2='External' Then
      EIM_Count[n][1]=EIM_Count[n][1]+1
      EIM_Scoring[n][1]=EIM_Scoring[n][1]+MCD→CURRENT_RECORD→
      EMetrics-Scoring
    Else If flag1='Non-Intrusive' and flag2='External' Then
      ENIM_Count[n][1]=ENIM_Count[n][1]+1
      ENIM_Scoring[n][1]=ENIM_Scoring[n][1]+MCD→CURRENT_RECORD→
      EMetrics-Scoring
    End If
    End If
    End If
    End If
  Else If CVSSMD→CURRENT_RECORD→Metrics_ID = 'TM008' Then
    If flag1='Intrusive' and flag2='Internal' Then
      IIM_Count[n][2]=IIM_Count[n][2]+1
      IIM_Scoring[n][2]=IIM_Scoring[n][2]+MCD→CURRENT_RECORD→
      EMetrics-Scoring
      Else If flag1='Non-Intrusive' and flag2='Internal' Then
        INIM_Count[n][2]=INIM_Count[n][2]+1
        INIM_Scoring[n][2]=INIM_Scoring[n][2]+MCD→CURRENT_RECORD→
        EMetrics-Scoring
      Else If flag1='Intrusive' and flag2='External' Then
        EIM_Count[n][2]=EIM_Count[n][2]+1
        EIM_Scoring[n][2]=EIM_Scoring[n][2]+MCD→CURRENT_RECORD→
        EMetrics-Scoring
      Else If flag1='Non-Intrusive' and flag2='External' Then
        ENIM_Count[n][2]=ENIM_Count[n][2]+1
        ENIM_Scoring[n][2]=ENIM_Scoring[n][2]+MCD→CURRENT_RECORD→
        EMetrics-Scoring
      End If
      End If
      End If
      End If
    Else If CVSSMD→CURRENT_RECORD→Metrics_ID = 'TM009' Then
      If flag1='Intrusive' and flag2='Internal' Then
        IIUM_Count[n][0]=IIUM_Count[n][0]+1
        IIUM_Scoring[n][0]=IIUM_Scoring[n][0]+MCD→CURRENT_RECORD→
        EMetrics-Scoring
        Else If flag1='Non-Intrusive' and flag2='Internal' Then
          INIUM_Count[n][0]=INIUM_Count[n][0]+1
          INIUM_Scoring[n][0]=INIUM_Scoring[n][0]+MCD→CURRENT_RECORD→
          EMetrics-Scoring
        Else If flag1='Intrusive' and flag2='External' Then
          EIUM_Count[n][0]=EIUM_Count[n][0]+1
          EIUM_Scoring[n][0]=EIUM_Scoring[n][0]+MCD→CURRENT_RECORD→
          EMetrics-Scoring
        Else If flag1='Non-Intrusive' and flag2='External' Then
          ENIUM_Count[n][0]=ENIUM_Count[n][0]+1

```

```

ENIUM_Scoring[n][0]=ENIUM_Scoring[n][0]+MCD→CURRENT_RECORD→
EMetrics-Scoring
  End If
  End If
  End If
  End If
  Else If CVSSMD→CURRENT_RECORD→Metrics_ID = 'TM010' Then
    If flag1='Intrusive' and flag2='Internal' Then
      IIUM_Count[n][1]=IIUM_Count[n][1]+1
      IIUM_Scoring[n][1]=IIUM_Scoring[n][1]+MCD→CURRENT_RECORD→
      EMetrics-Scoring
      Else If flag1='Non-Intrusive' and flag2='Internal' Then
        INIUM_Count[n][1]=INIUM_Count[n][1]+1
        INIUM_Scoring[n][1]=INIUM_Scoring[n][1]+MCD→CURRENT_RECORD→
        EMetrics-Scoring
      Else If flag1='Intrusive' and flag2='External' Then
        EIUM_Count[n][1]=EIUM_Count[n][1]+1
        EIUM_Scoring[n][1]=EIUM_Scoring[n][1]+MCD→CURRENT_RECORD→
        EMetrics-Scoring
      Else If flag1='Non-Intrusive' and flag2='External' Then
        ENIUM_Count[n][1]=ENIUM_Count[n][1]+1
        ENIUM_Scoring[n][1]=ENIUM_Scoring[n][1]+MCD→CURRENT_RECORD→
        EMetrics-Scoring
      End If
      End If
      End If
      End If
    End If
    End If
    End If
    End If
    Exit Loop
    Else
      Go To CVSSMD→NEXT_RECORD
    Loop
    End If
  End Do
  Step 19 Exit Loop
  Else
    Go To MCD→NEXT_RECORD
  Loop
  End If
  End Do
  Else
    Go To ASMCD→NEXT_RECORD
  Loop
  End If
Step 20 End Do
Step 21 n = n +1
Step 22 Go To VD→NEXT_RECORD
  Loop
Step 23 End Do
Step 24 Close All Databases
Step 25 Release All Memories Allocated
Step 26 Apply MCOQR (Misuse Case Oriented Quality Requirements)Metrics
to calculate and populate the Misuse Case Modeling Tree Worksheet
Step 27 Final Scoring will reveal the predicted Degree of Security of the
proposed application as per the following ranking:

```

3 Implementation Mechanism

On the basis of the data available in the central repository, a proposed algorithm will be used to retrieve, calculate, and fill in the worksheet Misuse Case Modeling. The proposed Misuse Case Oriented Quality Requirements (MCOQR) metrics will be applied to these populated data in order to derive the final data containing predicted counts of misuse cases, scoring and rankings revealing interrelated multidimensional levels of threat source, threat impact, counter measurement level, and dominant vulnerability type.

The final outcome will be in the form of a worksheet showing various levels of indicators and estimators in varied colors (for identification purpose) which may be used by the security requirements engineering team to analyze and interpret the level of security implementation in the software application well before its design and development. This shall also enable the team to further strengthen the security aspect of software by removing the defects of misuse case modeling with the introduction of more countermeasures.

4 Results and Validation

This proposed algorithm was applied to an industry real-life project (identity is hidden at the company's request), and the final result of the security assessment is calculated in accordance with the prescribed implementation mechanism. The level of security assurance is then compared to the security assurance of the other project, which did not apply the proposed algorithm. The study shows that the risk level is reduced by up to 40.5%. We do not provide the details of the validation results in this paper because of the page limit restriction; we will discuss them in our next paper.

5 Conclusion and Future Work

Our proposed work provides identification and classification of risk-based security requirements engineering and derived software security metrics with a sound step-by-step implementation mechanism which fits smoothly and easily into most of the software development lifecycle methodology. The proposed algorithm provides count and ranking for individual vulnerabilities, and a proper analysis may help the security expert to draft a comprehensive security requirement during the initial phase of software development process. Future work may include presentation and discussion of the proposed algorithm on a large scale of dataset of validation purpose.

References

1. Banerjee, C., & Pandey, S. K. (2009). Software security rules. *SDLC Perspective. arXiv preprint*.
2. Banerjee, C., Banerjee, A., & Murarka, P. D. (2014). Evaluating the relevance of prevailing software metrics to address issue of security implementation in SDLC. *International Journal of Advanced Studies in Computers, Science and Engineering*, 3(3), 18.
3. Banerjee, C., Banerjee, A., & Pandey, S. K. (2016). MCOQR (misuse case-oriented quality requirements) metrics framework. In *Problem Solving and Uncertainty Modeling through Optimization and Soft Computing Applications* (pp. 184–209). IGI Global.
4. Fenton, N., & Bieman, J. (2014). *Software metrics: A rigorous and practical approach*. CRC Press.
5. McGraw, G. (2006). *Software security: Building security in* (Vol. 1). Addison-Wesley Professional.
6. Mellado, D., Fernández-Medina, E., & Piattini, M. (2010, August). A comparison of software design security metrics. In *Proceedings of the Fourth European Conference on Software Architecture: Companion Volume* (pp. 236–242). ACM.
7. Pressman, R. S. (2005). *Software engineering: A practitioner's approach*. Palgrave Macmillan.
8. Robinson, S., Brooks, R., Kotidas, K., & Van Der Zee, D. J. (Eds). (2010). *Conceptual modeling for discrete-event simulation*. CRC Press.
9. Franceschini, F. (2010). *Advanced quality function deployment*. CRC Press.

Big Data Analytics for Data Quality Improvement to Enhance Evidence-Based Health Care in Developing Countries



Billy Mathias Kalema and Viola Vivian Busobozi

Abstract Organizations need to overcome all data analytics barriers in order to realize value from data by transforming it into insight leading to action that can add value to their businesses. Some of these barriers are as a result of the existing applications and technologies that are too rigid and have not been revised to match the increasing users' demands, whereas others arise from weaknesses in culture, stewardship, and governance that become salient when the need for quality data for decision-making increases. Attaining and keeping data quality is one of the most puzzling governance issues within organizations causing data quality errors that impede decision-making. Many health institutions in developing countries, like those in Africa, are faced with various challenges such as limited resources including manpower, unshared information, lack of privacy, poor drawing of insights from a variety of structured and unstructured data and insufficient budgets. The objective of this research work is to report on how big data analytics could be leveraged by these institutions to address these potential challenges. However, many health institutions still lack clarity of how big data analytics could be leveraged to improve data quality needed for evidence-based health care. This research work carried out a quantitative analysis of data collected from a health institution in South Africa. Results indicated that environment, tasks, data governance and structures, data quality management and technology are significant in improving data quality. This implies that to achieve and maintain data quality organizations need to pay attention to the elements such as people, process, technologies, and best practices that drive data governance. This study contributes to the ongoing debate on using big data analytics to enhance data quality.

Keywords Big data analytics · Data quality · Evidence-based health care · Data quality management · Data quality improvement

B. M. Kalema (✉) · V. V. Busobozi
Department of Informatics, Tshwane University of Technology, Pretoria, South Africa
e-mail: kalemabm@tut.ac.za

V. V. Busobozi
e-mail: busobozivv@tut.ac.za

© Springer Nature Singapore Pte Ltd. 2020
M. Pant et al. (eds.), *Performance Management of Integrated Systems and its Applications in Software Engineering*, Asset Analytics,
https://doi.org/10.1007/978-981-13-8253-6_4

1 Introduction

Organizations need to overcome all data analytics barriers in order to realize value from data by transforming it into insight leading to action that can add value to their businesses. Some of these barriers are as a result of the existing applications and technologies that are too rigid and have not been revised to match the increasing users' demands, whereas others arise from weaknesses in culture, stewardship, and governance that become salient when the need for quality data for decision-making increases [1]. Much as this is so, managing data quality is among the most vexing information management challenges, and as a result, every data design, be it logical or technical, should concern itself with various issues including but not limited to scalability, performance, adaptability, legacy, and package databases [2].

Attaining and keeping data quality is one of the most puzzling governance issues within organizations causing data quality errors that impede decision-making [2]. Additionally, Geiger [3] also noted that many organizations have persistent and long-standing data quality problems that they may not be known until a disaster happens. This implies that organizations need a proactive data quality management programme in place. Researcher [3] observed that data quality challenges may arise from various sources and from a number of issues since they are hidden and persistent and may remain unnoticed for a good period of time and may end up being propagated to other systems or business units as the connectivity increases.

Data quality can be improved through data governance more especially when business processes are streamlined; however, this may meet various challenges when received data is not properly analysed [4]. On the other hand, Chapman [2] indicates that at each stage of data governance, maximum care should be taken to avoid drawbacks and proper handling and analysis of received data should take precedence. From this understanding, Chapman [2] indicated these stages to include data capture and recording at the time of gathering, data manipulation prior to digitisation, identification of the collected data and its recording, digitisation of data, documentation of data, storage and archiving, presentation and dissemination as well as using the data.

The key aim of data governance is to define, approve, and communicate data strategies, policies, standards as well as describing business problems by intensively weighing the level of urgency and the value and quality data brings to the organization through analysis of the available data [4]. Researchers [4] also concur that data governance brings a standardized level of discipline to data management within an organization. However, they noted that traditional data governance practices are increasingly getting challenged by the new developments in data acquisition.

The new trends that have been singled out as challenges to traditional data quality include but not limited to the increasing use of agile methodologies for projects, big data, cloud deployment, and the adoption of self-service business intelligence (BI) and analytics [5]. This implies that to achieve and maintain data quality organizations need to pay attention to the elements such as people, process, technologies, and best practices that drive data governance.

1.1 *Big Data Analytics and Data Quality*

The advent of the big data era has seen the explosive growth of data in organizations and has challenged the ensuring of data quality, analysis, and mining of information and knowledge. Cai and Zhu [6] noted that poor data quality leads to low data utilization efficiency and brings serious decision-making mistakes. Data quality refers to the state of data fitting usage and conforms to user's requirements, goals, and objectives in a specific context [7, 8]. Data consistency and completeness are two central dimensions of data quality, whereby consistency refers to keeping data uniformly across the network and completeness refers to the degree to which all data necessary for current and future business activities are available in the data repository [9].

Maintaining consistency and completeness makes managing data quality (DQ) a worrisome information management issues in many organizations. Literature shows that business intelligence (BI) and analytics platforms have been immensely effective with structured data to augment business operations within organizations. However in the big data era, the BI challenges have changed dramatically, in terms of both goals and execution, whereby the simple processes of extract, transform, load (ETL) integration for structured enterprise data no longer meet the need [5, 9].

With big data, organizations are increasingly in need of using active data quality that deals with information related to the real world and things that are constantly moving outside their boundaries such as customers and their characteristics. This phenomenon has increased the generation of semi-structured and unstructured data within organization that complicates DQ control [10].

Big data is the exponential growth of data that is overwhelmingly big to fit the structures of the traditional relational database management systems (RDBMS) architectures in terms of transactional volumes, velocity responsiveness, and the quantity and or variety [11]. Big data makes scalability a challenge and complicates the making sense of unstructured data into actionable information [12].

According to Kwon et al. [9], the challenge of big data is less of the data volume than the quality of data needed for efficient delivery of services. Gandomi and Haider [10] alluded that if better analytics of big data are not put in place, the quality of data may be affected from four different perspectives, namely (a) failure to draw better insights from structured and unstructured data sets which affect the data sets from being used to address different user requirements and different outcomes, (b) confusing data as being with errors or having inconsistencies during migration from one source application to another, (c) lack of extension of historical data lifetime that could affect data governance, and (d) poor authenticity of information and lack of trust in data sources.

The big data era makes the volume of data to grow beyond the orders of magnitude, variety of data to evolve from the traditional structured datasets to unstructured data, and the increase in the velocity of data accumulation [11]. This revolution in the acquired data within an organization poses threats and new demands on data storage, analytical software, BI approaches to data governance and data quality [9, 12]. These challenges could be overcome if organizations boosts themselves with tools and

techniques needed to turn this data into insights and eventually value, such process is known as Big Data analytics [11].

This implies that much as data-driven is the modern mantra of business management, lack of analytics of big data may lead to multiple and complex chaotic situations within an organization and may impede on proper decision-making and impact on competitive advantage and or service delivery. Additionally, Cai and Zhu [6] identified four possible sources of challenges of data quality arising from big data.

These are (a) difficulty of data integration due to increased diversity of data sources, data types, and complex data structures, (b) difficulty to judge data quality within a reasonable amount of time due to increased data volume, (c) increased need of higher requirements for processing technology due to shortness of timeliness of data caused by rapid generation, (d) newness of research on data quality and lack of approved data quality standards in many countries.

Big data is continuously growing from terabytes to petabytes, zettabytes, and so on. Hence, the need for better analytics in any business processes that depend on data such as evidence-based health care is paramount for improving data quality to enhance decision-making [9]. There is a strong need for organizations to reduce the time and complexity involved in preparing data for analysis. Such a need is even more desired when dealing with a variety of data types and formats. At the same time, it is important to note that on its own data can be of little importance to an organization as it neither drive consensus nor inspire actions. This implies that organizations need to analyse their data and interpret its trends so as to transform it into something tremendously valuable.

1.2 Big Data Analytics, Data Quality, and Evidence-Based Health Care

Evidence-based medicine refers to the use of evidence from well-designed and well-conducted research to optimize decision-making in medicine [13]. Evidence-based medicine requires dependency on most recent research and relevant clinical data from a variety of sources while factoring in advanced analytics to improve patient care and outcomes. According to Baechle et al. [14], the increasing use of IT has seen healthcare data being generated from various unstructured sources including but not limited to clinical notes, digital devices, imaging, emails, laboratory tests, telematics, and third-party sources. They asserted that regardless of the sources, the quality of this data should be dependable in order to make meaningful decisions needed for diagnosis and treatment of patients. Cohen et al. [15] observed that due to the fact that most data needed for evidence-based is unstructured, and comes from various sources better analytics for this data is paramount.

As Mutale et al. [16] noted, health institutions in developing countries are faced with lots of challenges ranging from lack of resources including manpower, poor

access to information and facilities and all of these impact on the effectiveness of their evidence-based health care. However, with big data analytics access to massive amounts of structured and unstructured patient data could be obtained and analysed to support diagnosis of patients' conditions, match treatment with best outcomes, and predict patients at risk for disease or support hospital readmission [14].

2 Data Quality Theoretical Foundations

Researchers [17] observed that data quality issues such as duplication, missing information, formatting, and inaccurate profiling reflect in computational intelligence more especially when data are not preprocessed cleaned through cleansing, verification, formatting, and updating. They indicated that these issues could cause unwanted situations in evidence-based health care such as patients' distress, wastage of money, and increased organizational risks. Additionally, Chen et al. [7] indicate that data quality is influenced by technical, organizational, behavioural, and environmental factors. Their study reviewed 39 publications about data quality assessment in public health information system and revealed that data collection and use was given least attention. Cai and Zhu [6] identified the challenges of data quality and its assessment in big data era and established that data quality not only depends on its dimensions but also on business environment, processes, and users. They proposed a hierarchical data quality standard from the users' perspective which involves big data quality dimensions, quality characteristics, and quality indexes or indicators. The big data quality dimensions of availability, usability, reliability, relevance, and presentation quality were identified along with their elements that included accessibility, timeliness, authorization, credibility, definition/documentation, metadata, accuracy, consistency, integrity, completeness, auditability, fitness, readability, structure as well as indicators.

On the other hand, Espinosa and Armour [18] designed a framework for coordination and governance of big data analytics. Their study used the coordination theory and established that structural, operational, and relational practices are vital for big data analytics governance. Consequently, Alhassan et al. [19] analysed 31 peer-reviewed articles on data governance activities using the [20] data governance framework of five decision domains. Based on the content analysis method, their study established that eight major data governance action areas need to be empirically tested in future data governance research. These areas are data roles and responsibilities, policies, processes and procedures, standards, strategy, guidelines, technologies, and requirements.

To embrace the factors of data quality and big data analytics identified from the literature, this study utilized contingency theory [21], deferred theory of action [22] and the technology, organizational, and environment (TOE) model [23]. From the contingency theory, a management information system (MIS) contingency model's constructs were used [24]. Constructs used included strategy, structure, environment,

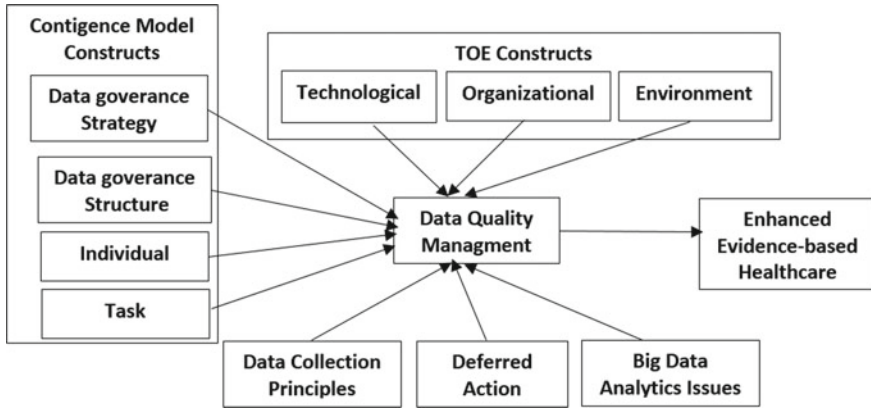


Fig. 1 Research model

technology, tasks, and individual. From contingency theory, one construct was used which was deferred action.

Data quality management, roles and responsibilities, policies, processes and procedures, standards, strategy, guidelines constructs were also used depending on their classifications. The TOE model explains the potential organizational shapes for initiating big data analytics use; hence, its three constructs of technology, organization, and environment were incorporated. The research model of this study is as demonstrated in Fig. 1.

2.1 Construct Description and Hypothesis Development

Based on the constructs of the research model, this study hypothesized the following relationship in relation to data quality management and evidence-based health care.

- (a) **Data governance strategy**—This construct focuses on the availability of strategy, policies, principles, procedures, data sets, data collection goals that must be in line with organization strategy. From this understanding, hypothesis (**H1**) was developed.
H1: Data governance strategy when mediated by data quality management enhances evidence-based health care.
- (b) **Data governance structure**—This is viewed in terms of decision-making authority on data quality issues which can either be centralized or decentralized. That is, who decides on what, when, and where about data principles, architecture, metadata, quality, and lifecycle. From this understanding, hypothesis (**H2**) was developed.
H2: Data governance structure when by data quality management enhances evidence-based health care.

- (c) **Individuals**—This looks at roles, responsibilities, accountability, and the way employees perceive data quality management to enhance evidence-based health care. Based on this understanding, hypothesis (**H3**) was developed.
H3: Individuals when mediated by data quality management enhance evidence-based health care.
- (d) **Tasks**—These are business processes needed for data quality activities that should be in line with data governance principles to achieve better data quality management. These tasks include but not limited to user requirements and organizational needs. This understanding led to the development of hypothesis (**H4**).
H4: Tasks when mediated by data quality management enhance evidence-based health care.
- (e) **Technological**—This looks at data technologies used to support the way people work, but not the way it works, thus making it easy to extract the needed information by the users. From the reviewed literature, hypothesis (**H5**) was developed.
H5: Technological issues when mediated by data quality management enhance evidence-based health care.
- (f) **Organizational**—This looks at the top management and their support to big data, organizations size, and the financial state of the organization such as finances, training of users, structure, top management support. This led to the development of hypothesis (**H6**).
H6: Organizational issues when mediated by data quality management enhance evidence-based health care.
- (g) **Environmental issues**—This construct focuses on the culture and norms, the ability to change to new innovations, different business unit's integration internally within the organization. Internal is how systems interact to ensure data quality and externally the policies on data quality. From this understanding, hypothesis (**H7**) was developed.
H7: Environmental issues when mediated by data quality management enhance evidence-based health care.
- (h) **Data collection principles**—This construct investigated the influence of data collection principles and processes on the management of data quality within organization. Data collection embraces definition, collection, processing, and representation, whereas at the same time it impacts data and information quality. Based on this understanding, hypothesis (**H8**) was derived.
H8: Data collection principles when mediated by data quality management enhance evidence-based health care.
- (i) **Deferred action**—This construct focused on future actions that can be done to improve data quality through big data analytics and data quality management. From the reviewed literature, hypothesis (**H9**) was developed.
H9: Deferred action when mediated by data quality management enhances evidence-based health care.
- (j) **Big data analytics issues**—This constructs investigated the big data quality dimensions of availability, usability, reliability, relevance, and presentation qual-

ity along with their elements of accessibility, timeliness, authorization, credibility, definition/documentation, metadata, accuracy, consistency, integrity, completeness, auditability, fitness, readability, structure. Based on this understanding, hypotheses (**H10** and **H11**) were developed.

H10: Big data quality dimensions when mediated by data quality management enhance evidence-based health care.

H11: Big data quality dimensions directly influence the enhancement of evidence-based health care.

- (k) **Data quality management**—This focuses at ensuring that high-quality data is achieved, and it embraces the factors needed for good data governance and improvement of standards. This construct mediated other factors as good data quality management leads to good data needed for decision-making. From this understanding, hypothesis (**H12**) was developed.

H12: Data quality management directly influences the enhancement of evidence-based health care.

3 Methodology

Based on the research model, a measuring instrument in form of a close-ended questionnaire was designed and used for data collection. In total, 200 questionnaires were distributed to employees of health institutions in South Africa. Of the distributed questionnaires, 163 were returned and 152 were usable. Collected questionnaire was screened, coded, and transcribed into SPSS v 22.0 for analysis. The coding of the constructs was done, whereby data governance strategy was coded as DGStr, data governance structure as DGStru, individual as IND, tasks remained unchanged, data collection principles as DCP, deferred action as DefA, big data analytics issues as BDAI, technological as TECH, organizational as ORG, environmental as ENV, data quality management as DQM, and enhanced evidence-based health care as EVBH.

The questionnaire was tested for reliability, and its overall coefficient was found to .794. The questionnaire constructs were also tested independently, and many were found to be above the recommended threshold of .7 [25]. Participants were sampled based on their closeness with data operations such as data collectors, data captures, decision-makers, and IT personnel.

4 Results

Upon the completion of reliability checks, correlation to establish how each construct relates to the other was conducted. The most significant values were established and these were found to be between data governance strategy and deferred action, data governance structures and individual characteristics, data quality management and enhancement of Evidence-based healthcare, individual characteristics with, Big Data

analytics issues, tasks, deferred action, data quality management and enhancement of Evidence-based healthcare. Environmental and technological factors also exhibited a good correlation as well as big data analytics issues with data quality management.

4.1 Regression Analysis

Further to correlations analysis, this study also carried out a multiple regression analysis to determine the contribution of each construct to the enhancement of evidence-based health care. Table 1 demonstrates the results of regression analysis.

Since data quality management entails the establishment and deployment of roles, responsibilities, policies, and procedures concerning the acquisition, maintenance, dissemination, and disposition of data, several constructs as illustrated in the model were tested to investigate their contribution. Results demonstrated in Table 1 show big data analytic issues (BDAI) significantly and contribute highly 53.1% ($\beta = .531$) to data quality management. This was followed by data governance structures (DGStru) with a contribution of 29.6%, data collection principles with 28.6%, and organizational factors with 26.6%. On the other hand, deferred action had the least insignificant contribution of 4.9% followed by data governance strategy with 10.6%.

The findings as demonstrated in Table 1 imply that several factors are needed to augmentation data besides technology to help in data quality management. And when these factors are leveraged, an organization is empowered to make decisions

Table 1 Multiple regression analysis results

| Model | | Unstandardized coefficients | | Standardized coefficients | t | Sig. | Collinearity statistics | |
|-------|------------|-----------------------------|----------------|---------------------------|--------|------|-------------------------|-------|
| | | B | Standard error | Beta | | | Tolerance | VIF |
| 1 | (Constant) | 3.283 | .546 | | 6.012 | .000 | | |
| | DGStra | .104 | .065 | .106 | 1.605 | .111 | .806 | 1.241 |
| | DGStru | .221 | .072 | .296 | 3.065 | .003 | .379 | 2.635 |
| | IND | -.122 | .061 | -.073 | -1.995 | .048 | .418 | 2.394 |
| | Tasks | -.182 | .064 | -.234 | -2.855 | .005 | .525 | 1.905 |
| | DCP | .205 | .059 | .286 | 3.472 | .001 | .520 | 1.924 |
| | DefA | -.043 | .069 | -.049 | -.627 | .531 | .570 | 1.754 |
| | BDAI | .408 | .078 | .531 | 5.237 | .000 | .422 | 2.369 |
| | TECH | -.173 | .085 | -.124 | -2.040 | .043 | .956 | 1.046 |
| | ORG | -.197 | .068 | -.266 | -2.901 | .014 | .564 | 1.773 |
| | ENVT | -.218 | .073 | -.182 | -2.996 | .003 | .962 | 1.040 |
| DQM | .376 | .075 | .470 | 5.040 | .000 | .406 | 2.464 | |

a. Dependent variable: EVBH

with the confidence and accuracy accrued from data quality. Similarly, it is evident from these findings that data quality management has expanded to incorporate more than just relational data stores, and the analytics need to go beyond the traditional process to include tools that give tremendous value to everyday business users and such makes big data analytics an important factor.

Table 1 also demonstrates the results of the measure of multicollinearity using the variance inflation factor (VIF) and tolerance. Results indicate that all values of VIF were below the recommended threshold of 10 [26]. Existence of multicollinearity could threaten the statistical significance of independent variables. Belsley [27] alludes that the lack of multicollinearity indicates that the independent constructs in the regression model predict the dependent variable unconventionally without the influence of one another.

4.2 Hypothesis Testing

This study suggested twelve hypotheses in relation to data quality management enhancing evidence-based health care. Table 2 demonstrates the results of the tested hypotheses showing their significance at $p = .05$.

5 Discussion, Conclusion, and Recommendations

Results demonstrated in both Tables 1 and 2 indicate that there are many factors that need to be considered for big data analytics to improve data quality management needed to enhance evidence-based health care in developing countries. Data quality management needed for effective decision-making and evidence-based health care is based on sound decision-making. Out of the suggested 12 hypotheses, 10 were accepted while only two were rejected.

Of the rejected hypotheses (H1) and (H9), it could be linked to the fact that as the organization's data management, ecosystem continues to evolve which is the case with the advent of big data, it is extremely important that the existing data governance practices also rapidly evolve, so that organizations can remain competitive. This could imply that data governance policies, rules, and strategies need to constantly be changed in order for them to remain aligned with business goals; otherwise, they become irrelevant [2]. More still, as the conventional approach to data quality promises oversight and control over the input, big data on the other hand may threaten the quality of the outcome.

The findings of this study are in agreement with those of [18] who argued that structural, operational, and relational practices are vital for big data analytics governance, and once the overall data governance goals are conceptualized, then organizations need to implement technologies to support those goals and objectives. This implies that in the changing data governance and management environment, strategies need

Table 2 Hypothesis testing

| Hypotheses | <i>P</i> -value | Outcome |
|--|------------------|----------|
| H1: Data governance strategy when mediated by data quality management enhances evidence-based health care | $P = .111 > .05$ | Rejected |
| H2: Data governance structure when by data quality management enhances evidence-based health care | $P = .003 < .05$ | Accepted |
| H3: Individuals when mediated by data quality management enhance evidence-based health care | $P = .048 < .05$ | Accepted |
| H4: Tasks when mediated by data quality management enhance evidence-based health care | $P = .005 < .05$ | Accepted |
| H5: Technological issues when mediated by data quality management enhance evidence-based health care | $P = .043 < .05$ | Accepted |
| H6: Organizational issues when mediated by data quality management enhance evidence-based health care | $P = .014 < .05$ | Accepted |
| H7: Environmental issues when mediated by data quality management enhance evidence-based health care | $P = .003 < .05$ | Accepted |
| H8: Data collection principles when mediated by data quality management enhance evidence-based health care | $P = .001 < .05$ | Accepted |
| H9: Deferred action when mediated by data quality management enhances evidence-based health care | $P = .531 > .05$ | Rejected |
| H10: Big data quality dimensions when mediated by data quality management enhance evidence-based health care | $P = .000 < .05$ | Accepted |
| H11: Big data quality dimensions directly influence the enhancement of evidence-based health care | $P = .024 < .05$ | Accepted |
| H12: Data quality management directly influences the enhancement of evidence-based health care | $P = .000 < .05$ | Accepted |

not to be static rather change from time to time to fit business requirements. Another argument is that quality of data exists solely in the eyes of the customer depending on the value they perceive with respect to meeting their needs. Since customers' needs may vary from time to time so do the need to change strategies to meet these changing requirements.

The acceptance of the 10 hypotheses emphasizes the fact that organizations need data quality management that combines business-driven and technical perspectives to respond to strategic and operational challenges demanding high-quality corporate data. Such may include but not limited to paying attention to data collection, organization characteristics, storage, processing, as well as technical aspects needed to present data. This also implies that organizations need to document any process or request relating to data sources, have in place standards, processes and tools needed to handle and control data protection as well as having capacity of controlling access to data including enforcing log user activity, and handling of data that is in many instances duplicated.

Several others researchers such as [2, 6, 8, 9, 15] also noted that total data quality management (TDQM) which is a key aspect of attaining quality data can only be

achieved if all factors needed for continuous defining, measuring, analysing, and improving information quality are put into consideration. However, data should have a purpose and value to the extent that even if it is accurate and has inherent quality, but without value to the enterprise such data should be considered useless.

5.1 Recommendations and Future Research

Many organisations realize that an integrated approach to data management has a positive impact throughout the business ranging from accelerating decision-making to improving efficiency and effectiveness in the provision of services. This study therefore recommends that organizations should use principles which are core beliefs that create a link between policies, processes, and behaviours for information asset management.

5.2 Contribution to the Study

Theoretically, this study contributes to the ongoing debate of the role of big data analytics in enhancing data quality and data quality management for better decision-making. This study has highlighted factors that need to be paid attention to improve data quality management. This framework could be based on by other researchers to extend research in this domain, and by so doing, this study will be making a significant theoretical contribution in the computing domain. On the other hand, the findings of this study will assist leaders within organizations to take proactive approach to data and understand the impact data can have on decisions, planning, and forecasting. By so doing, this study will be contributing practically to the organizational management.

5.3 Conclusions

With the emerging computing trends of big data analytics, there is a need for data quality management practices to be established in organizations. Big data results in business intelligence and analytics. Due to inaccurate data and inability to consolidate data from business units, organizations cannot generate meaningful business intelligence. More so unstructured data make it hard to be mined. Data whether structured, unstructured, or semi-structured should be analysed with the purpose of fitness for use. Therefore, quality of data is assumed if the analysed data is aligned to meet the user's objectives, goals and in a specific context. Management of data quality should be done in such way that the data conforms to the people's needs, operational procedures, policies, and processes.

References

1. Loshin, D. (2010). *Evaluating the business impact of poor data quality*. Knowledge integrity incorporated. Business Intelligence solutions.
2. Chapman, A. D. (2005). *Principles of data quality*, version 1.0. Report for the Global Biodiversity Information Facility, Copenhagen.
3. Geiger, J. G. (2004). The most critical initiative you can implement. *Data warehousing management and quality*, Paper 098-129. Boulder, Co: Intelligent Solutions, Inc.
4. Capiello, C., Caro, A. Rodriguez, A. & Caballero, I. (2013). An approach to design business processes: Addressing data quality issues. In *Proceedings of the 21st European Conference on Information Systems*, 6–8 June.
5. Michael, K., & Miller, K. W. (2013). Big data: New opportunities and new challenges. *IEEE Computer Society*, 46(6), 22–24.
6. Cai, L., & Zhu, Y. (2015). The challenges of data quality and data quality assessment in the big data era. *Data Science Journal*, 14(2).
7. Chen, H., Hailey, D., Wang, N., & Yu, P. (2014). A review of data quality assessment methods for public health information system. *International Journal of Environmental Research and Public Health*, 11(5).
8. Mustafa, A., Ramin, K., & Phillip, T. (2016). Metadata-based data quality assessment. *VINE Journal of Information and Knowledge Management Systems*, 46(2), 232–250.
9. Kwon, O., Lee, N., & Shin, B. (2014). Data quality management, data usage experience and acquisition intention of big data analytics. *International Journal of Information Management*, 34, 387–394.
10. Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, 35(2), 137–144.
11. Kalema, B. M., & Mokgadi, M. (2017). Developing countries organizations' readiness for big data analytics. *Problems and Perspectives in Management*, 15(1–1), 260–270.
12. Hilbert, M., & López, P. (2011). The world's technological capacity to store. *Communications in Computer and Information Science*, 332, 60–65.
13. Evidence-Based Working Group. (1992). Evidence-based medicine. A new approach to teaching the practice of medicine. *JAMA* 268(17), 2420–2425.
14. Baechle, C., Agarwal, A., & Zhu, X. (2017). Big data driven co-occurring evidence discovery in chronic obstructive pulmonary disease patients. *Journal of Big Data*, 4(9), 1–18.
15. Cohen, A. M., Adams, C. E., Davis, J. M., Yu, C., Yu, P. S., Meng, W., et al. (2010). The essential role of systematic reviews, and the need for automated text mining tools (pp. 376–380).
16. Mutale, W., Chintu, N., Amoroso, C., Awoonor-Williams, K., Phillips, J., Baynes, C., et al. (2013). Improving health information systems for decision making across five sub-Saharan African countries: Implementation strategies from the African Health Initiative. *BMC Health Services Research*, 13(9), 12.
17. Nahar, J., Imam, T., Kevin, Tickle, S., & Garcia-alonso, D. (2013). Issues of data governance associated with data mining in medical research: Experiences from an empirical study. In E. J. S. Hovenga & H. Grain (Eds.), *Health information governance in a digital environment*. IOS Press.
18. Espinosa, A. J., & Armour, F. (2016). The big data analytics gold rush: A research framework for coordination and governance. In *49th Hawaii International Conference on System Sciences*, 5–8 Jan.
19. Alhassan, I., Sammon, D., & Daly, M. (2016). Data governance activities: An analysis of the literature. *Journal of Decision Systems*, 25(1), 64–75.
20. Khatri, V., & Brown, C. V. (2010). Designing data governance. *Communications of the ACM*, 53(1).
21. Vroom, V. H., & Yetton, P. W. (1973). *Leadership and decision-making*. Pittsburgh: University of Pittsburgh Press.
22. Patel, N. V. (2006). *Organisation and systems design: Theory of deferred action* (p. 253). Palgrave Macmillan.

23. Tornatzky, L. G., & Fleischer, M. (1990). *The processes of technological innovation*. Lexington, MA: Lexington Books.
24. Weills, P., & Oslon, M. H. (1989). An assessment of contingency theory of management information systems. *Journal of management information system (JMIS)*, 6(1), 59–86.
25. Pallant, J. (2010). *SPSS survival manual: A step by step guide to data analysis using SPSS for Windows*. New York: McGraw-Hill.
26. Aiken, L. S., & West, L. G. (1991). *Multiple regression: Testing and interpreting interactions*. Newbury Park, CA: Sage.
27. Belsley, D. (1991). *Conditioning diagnostics: Collinearity and weak data in regression*. New York: Wiley.

Securing the Data Deduplication to Improve the Performance of Systems in the Cloud Infrastructure



Nishant N. Pachpor and Prakash S. Prasad

Abstract Data duplication is a data quality problem which may exist in database system where the same record is stored multiple times in the same or different database systems. Data duplication issue may lead to issues like data redundancy, wasted cost, lost income, negative impact on response rate, ROI, and brand reputation, poor customer service, inefficiency and lack of productivity, decreased user adoption, inaccurate reporting, less informed decisions, and poor business process. The solution to the problem of data duplication may be countered with data deduplication which is often termed as intelligent compression or single instance storage. Data deduplication eradicates duplicate copies of information resulting in the reduction of storage overheads and in enhancement of various performance parameters. The recent study on data deduplication has shown that there exists modern data redundancy in primary storage in the cloud infrastructure. Data redundancy can be reduced in primary storage system of cloud architecture using data deduplication. The research work carried out highlights the identified and established methods of data deduplication based on capacity and performance parameters. In the research work, the authors have proposed a performance-oriented data (POD) deduplication scheme which improves performance and primary storage system in the cloud. In addition to this, security analysis using encryption technique has also been performed and demonstrated to protect the sensitive data after the completion of deduplication process.

Keywords Data deduplication · Security · Cloud · Storage capacity · Compression · Decompression

N. N. Pachpor (✉) · P. S. Prasad
P.I.E.T, Nagpur, India
e-mail: nishantpachpor@gmail.com

P. S. Prasad
e-mail: praksahsprasad@gmail.com

© Springer Nature Singapore Pte Ltd. 2020
M. Pant et al. (eds.), *Performance Management of Integrated Systems and its Applications in Software Engineering*, Asset Analytics,
https://doi.org/10.1007/978-981-13-8253-6_5

1 Introduction

In any industry, the employees copy and transfer data from one storage location to other storage location on daily basis. In this process, most of the time, same type of data is stored on the system resulting in redundancy. Consider a real-world example in an IT industry; the management sends an email to all employers with an attachment of 5 MB file which may be in any format like text, doc, and pdf. If everyone backs up the attachment of that 5 MB on the same server, the server will contain multiple copies of that 5 MB attachment on different locations and that is called data duplication.

Data deduplication is a specialized technique of removing the redundant chunk of data in the file. It helps in saving storage space. In other words, it can be said that by means of deduplication the user may ensure that unique instant of data is stored on the primary storage system in the cloud. As per the available information these days, there are a number of the techniques which may be used for the process of data deduplication. Data deduplication may be categorized in two ways, i.e., source-based and target-based.

In Fig. 1, a simple classification of data deduplication is shown. In source-based data deduplication, the redundant data is removed from data blocks before transmitting the data on client as well as server level. In source-based data deduplication, there is no need of additional hardware and increase in bandwidth and storage helps in improving the performance of the system.

In target-based data deduplication, storage data is transmitted across the network or in a remote location. Target-based data deduplication increases the cost compared to source-based deduplication. Target-based data deduplication is further subclassified into two techniques, i.e., inline and post-processing. In inline deduplication, the duplicate data gets deleted while writing data resulting in reduced amount of space storage requirement. In post-processing deduplication, duplicate data is checked and removed after the data is written in the file and then the data gets stored in the disk.

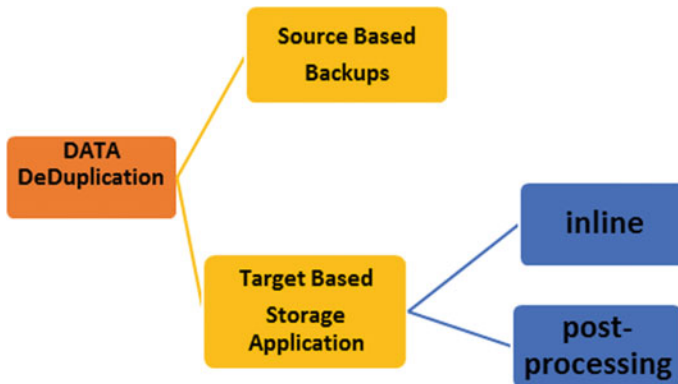


Fig. 1 A simple relation between these factors



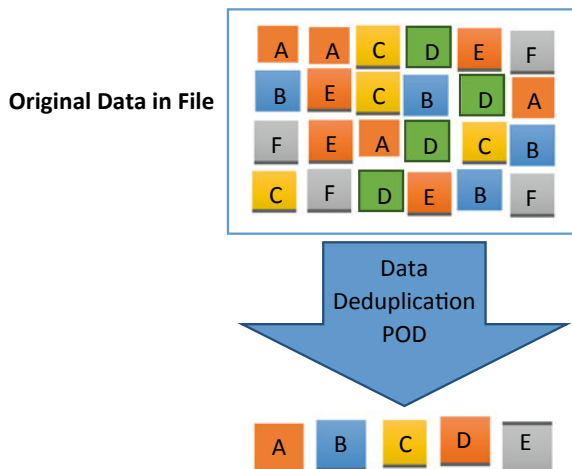
In recent study carried out, explosive nature and volume of data along with input/output bottleneck issue have become increasingly challenging for big data [1, 2] having concerns about the performance of the system because huge capacity of duplicate data has been found on cloud. It was also found that the volume of duplicate data increased by 7 ZB per year. It is estimated that in next ten years it will reach up to 35 ZB.

Data deduplication scheme such as iDedup [3] and offline dedup [4] which are known for increasing the performance of primary storage, it was found that they are not checking the small-size file which are stored in the cloud and as such the system performance of the system and network is getting affected. For improving the data capacity, the performance-oriented data deduplication technique also called POD may be used. POD employs two new technologies for improvement of primary storage in cloud storage, i.e., selective dedup and iCache which is a memory management scheme. The above two mentioned technologies provide easy access to the memory and reduce the traffic on network during data transfer and reading and writing data in the system.

In Fig. 2, it is shown that same file(s), for example, attachment in an email probably downloaded by a number of employees of an organization is stored in different locations on the organization’s storage server for future use or reference. After data deduplication method(s) is used to remove the duplicate data from data block, an increase in the bandwidth for data transfer can be observed over a network.

In Fig. 2, we propose a performance-oriented I/O deduplication method which checks the performance-oriented rather than a capacity-oriented I/O deduplication. In the POD, removing duplicate data reduces the file size which helps to overcome the traffic on a network while transferring data and reading–writing data. While using the iDedup which is a capacity oriented data deduplication. It checks only large size file and skip the small size file data like is 4–10 kb and there may be some amount of data duplication.

Fig. 2 Data deduplication



For improving performance of the primary storage and Input-output performance of a system in the cloud storage. It will check first the data in file before saving cloud storage. By using performance-oriented data deduplication (POD) method which removes duplicate data from capacity oriented and small size file. Performance-oriented data deduplication can be further subclassified into two types, i.e., selective dedup and iCache. For reducing the performance of data deduplication, selective dedup employs the request-based techniques. In iCache data deduplication technique which uses by using memory management scheme. In memory management scheme the fragmentation of data which helps to improve the performance of primary storage systems. iCache technique is suitable for bursty read traffic and the bursty write traffic on network.

2 Literature Survey

iDedup and offline dedup, these two existing technologies are used to check the data deduplication in the system for the improvement of primary storage in the cloud. These methods primarily focus on testing the only capacity-based data or file, and it selects the large size and bypasses the small-size file request on the cloud. The small I/O applications only account for a small fraction of the storage capacity requirement, which makes deduplication unprofitable and potentially counterproductive in view of the significant overhead deduplication.

Here, we observe that the small files lead to issues in primary storage systems and affect the performance of the systems. While data travel on network, the small size file exhibit I/O buses and affect the performance of primary storage [1].

2.1 *Disadvantages of Existing System*

The existing data deduplication methods only work for capacity-based file check rather than small file. It cannot check the performance of the system and does not provide any argument that why small-size files are not checked. The critical issue of the system is primary storage.

The existing data deduplication methods are directly applied to the data or file which causes space contention on storage system. Data deduplication introduces the significant index-memory management scheme overhead to the existing system. Memory management scheme split into multiple small data chunks that are often located in non-sequential locations on disks after deduplication. This fragmentation of data can cause a subsequent read request to invoke many, often random, drive I/O operations, leading to performance degradation.

In the existing system for data deduplication scheme for primary storage like iDedup and dedup, they only work for a large capacity of data duplication but not provide any security of data.

Table 1 Comparison between POD and other techniques

| Features | Capacity | Performance | Small I/O | Large I/O | Scope of improvement {security} |
|-------------------|----------|-------------|-----------|-----------|---------------------------------|
| I/O dedup [6] | No | Yes | No | No | No |
| iDedup [7] | Yes | No | No | Yes | No |
| Offline dedup [3] | Yes | No | No | Yes | No |
| POD [2] | Yes | Yes | Yes | Yes | Yes |

Table 1 shows the comparison of the different methods of data deduplication with various features like capacity, performance, small I/O, and large I/O. In POD prototype, we have the scope for improvement, i.e., implementation of security of data after the data deduplication. By applying Encryption and Decryption technique to ensure that data file will secure in primary storage as well as on cloud [2, 5] (Table 2).

3 Proposed System

The main problem in primary cloud storage is performance, and we propose a performance-oriented data deduplication scheme called POD. Instead of a capacity-oriented one (e.g., iDedup), the I/O performance of primary storage systems in the cloud can be improved by taking into account the characteristics of the workload. POD takes two types of approach, i.e., selective dedup and iCache. For reducing data deduplication, selective dedup checks the request-based techniques. In iCache by using memory management scheme fragmentation of data which helps to improve the performance of primary storage systems. iCache technique is suitable for bursty read traffic and the bursty write traffic.

Advantages of Proposed System:

1. The POD technique improves the performance and saves capacity of primary storage systems in the cloud
2. After applying POD prototype, we also do the encryption and decryption of data in the cloud that helps to secure the data as well as reduce redundancy of data.

Design Objective of Architecture is

1. Reducing the traffic for reading and writing the data in the primary storage.
2. Improving the performance of storage system.
3. Providing the data deduplication POD technique to reduce the data duplication.
4. Providing encryption and decryption security to secure the data inside the file.
5. With the help of POD and security, we can store the data in the cloud.

Table 2 Related work

| Sr.No. | Methodology | Advantage | Limitation | Future scope |
|--------|--|---|--|--|
| [1] | Performance-oriented I/O deduplication [1, 2] | <ul style="list-style-type: none"> • Primary storage performance improved • Capacity-based file or data checking • Easier to reading and writing data on network | Authentication and authorization of data | <ul style="list-style-type: none"> • To provide an authentication of system • Reduce the file by using encryption of file • Providing key, only owner can open the file |
| [2] | In cluster file system block-level data deduplication [6] | <ul style="list-style-type: none"> • File removes the duplication up to 80% | Skip the small size of file, and only check capacity-oriented data or file check | |
| [6] | Capacity-oriented inline data deduplication [4] | <ul style="list-style-type: none"> • iDedup • Workloads, while minimizing extra IOs | Bypass all the small-size requests, authentication of data | |
| [5] | Compression approach and a uniform chunk size distribution [8] | <ul style="list-style-type: none"> • Offline dedup • The Windows server operating system | | |
| [7] | Deduplication-aware resemblance detection and elimination scheme [9] | DARE only consumes about 1/4 and 1/2, respectively, of the computation and indexing | Security of data | |

Algorithm:

1. For checking the file duplicacy, upload file, search the duplicate file on cloud, remove the duplicate file, and give the feedback to user. If not found duplicate file then give permission to stored file in cloud storage.
2. Read file from cloud.
3. Check the duplicate data in the file.
4. If duplicate data found in file, inform the user and give the feedback.
5. Generate the security key.
6. Apply the encryption to file.
7. Generate key for decryption. Apply the decryption process.

4 System Architecture

Figure 3 shows the system architecture of POD. The POD is different for accessing data reading and writing from cloud and HDD.

As stated earlier, the POD has two methods which remove the data deduplication, i.e., selective dedup and iCache. By using the selective dedup, we can split the data into data blocks for writing data in memory. Checking the data blocks for presence of redundancy data in file. If redundancy is found, the duplicate data is removed from chunks of block and store to primary storage. iCache monitors the request and checks the hit request for reading and writing. Both components help to remove the large amount of data deduplication which improve the performance of system.

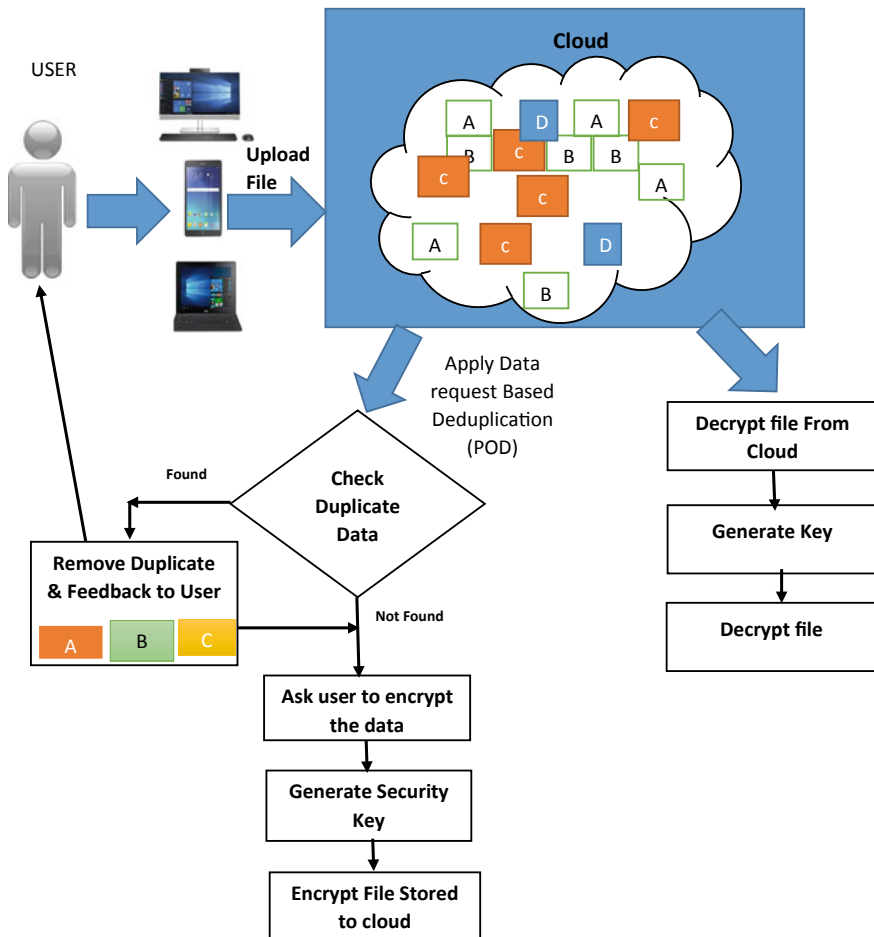


Fig. 3 System architecture of data deduplication



Through the encryption, we ensure the data is secure in the storage. When the data owner challenges to checking the data from cloud with its eligibility of data ownerships and issue the re-encryption key that can convert encrypted data into decrypted data.

In system architecture of data deduplication technique user can upload file from anywhere to the cloud as shown in Fig. 3. User can be uploading file from LAN/MAN/WAN, the proposed system first checks whether any duplicate file exists on cloud and if found it removes the file and gives feedback to the file owner. If duplicate file is not found, then the data inside the file is checked for redundancy, and if it is found the data is removed. Create a security key for the encryption and decryption of file, and ask the user for security of file for applying the process of encryption; if user wants to re-read the data, then the proposed system generates decryption key and applies the process of decryption [1, 2].

5 Result and Snapshots of Project

This section discusses the results and outcome of the proposed work. Snapshot of the software developed is also shown as follows:

5.1 Login Page

Login page is used for providing seven different types of access levels to different users accessing the system (Fig. 4).

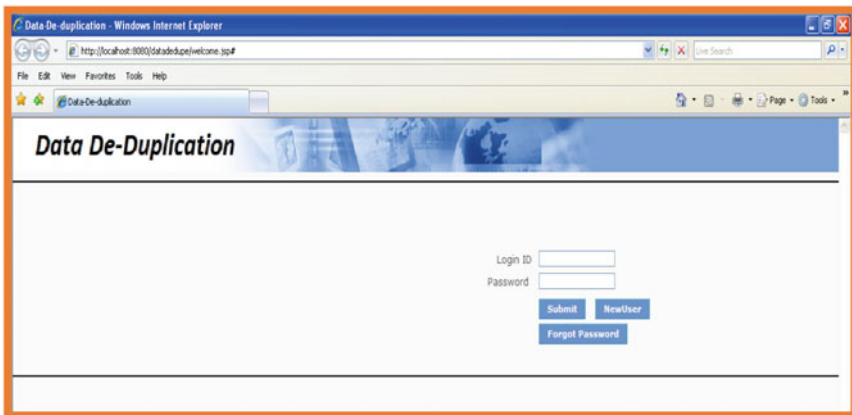
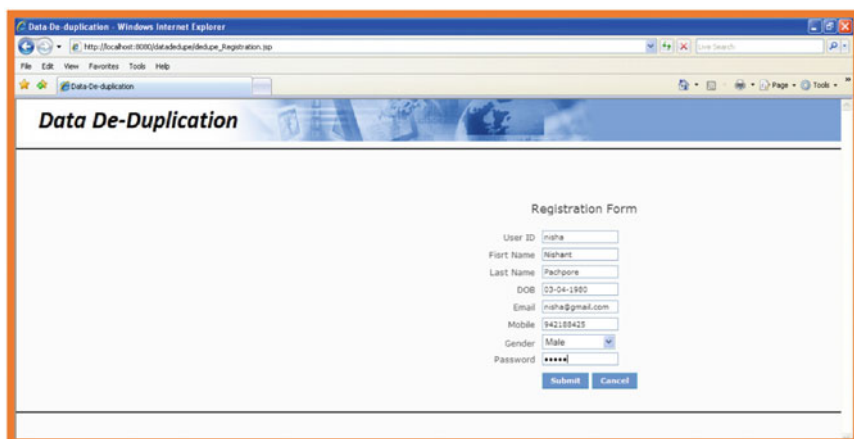


Fig. 4 Login page

5.2 Registration

See Fig. 5.



The screenshot shows a web browser window titled "Data De-duplication - Windows Internet Explorer". The address bar shows the URL "http://localhost:8080/DataDeDuplication/Registration.jsp". The page header features the text "Data De-Duplication" in a blue banner. Below the banner, the page is titled "Registration Form". The form contains the following fields and values:

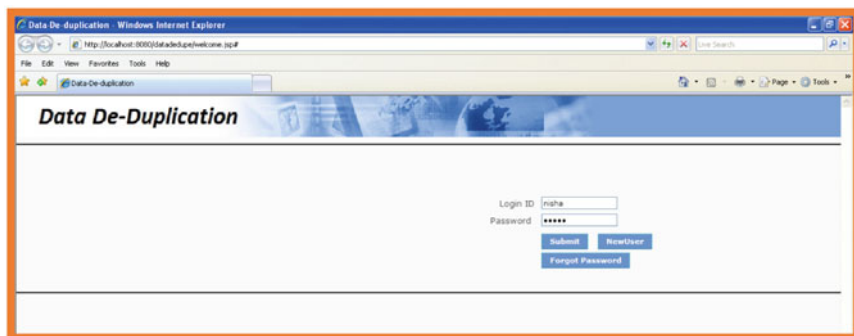
| | |
|------------|-----------------|
| User ID | nisha |
| First Name | Nishant |
| Last Name | Pachpore |
| DOB | 03-04-1990 |
| Email | nisha@gmail.com |
| Mobile | 942188425 |
| Gender | Male |
| Password | ***** |

At the bottom of the form, there are two buttons: "Submit" and "Cancel".

Fig. 5 Registration page

5.3 Login with Register User

Login user successfully render on functional page which has the functionality like file upload, encrypt/decrypt, file duplicate, data duplicate (Figs. 6 and 7).



The screenshot shows a web browser window titled "Data De-duplication - Windows Internet Explorer". The address bar shows the URL "http://localhost:8080/DataDeDuplication/welcome.jsp". The page header features the text "Data De-Duplication" in a blue banner. Below the banner, the page is titled "Login with Register User". The form contains the following fields and values:

| | |
|----------|-------|
| Login ID | nisha |
| Password | ***** |

At the bottom of the form, there are three buttons: "Submit", "New User", and "Forgot Password".

Fig. 6 Login with register user

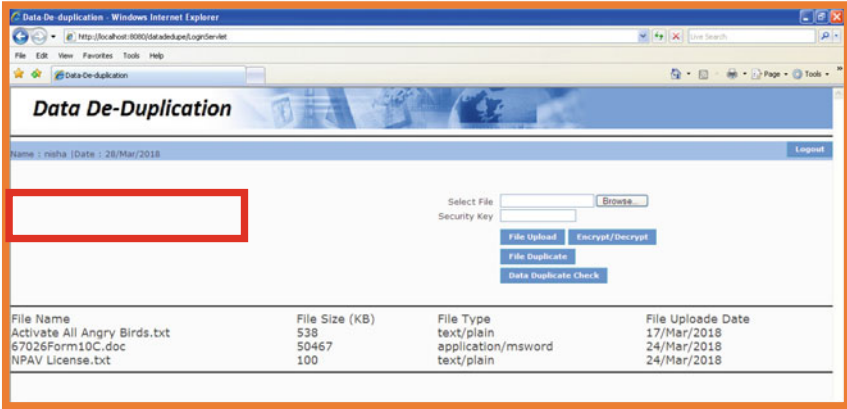


Fig. 7 User logs in successfully

5.4 File Upload Functionality

- (1) Select a file to upload on cloud through browser button, and click on file upload button (Fig. 8).

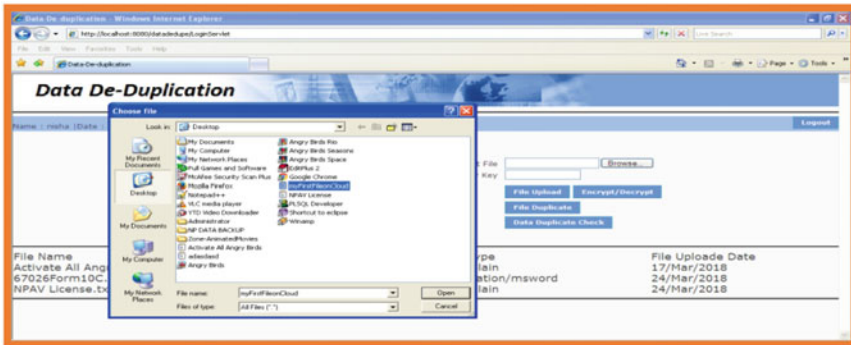


Fig. 8 File upload functionality

- (2) Select file to upload on cloud and upload (Figs. 9 and 10).

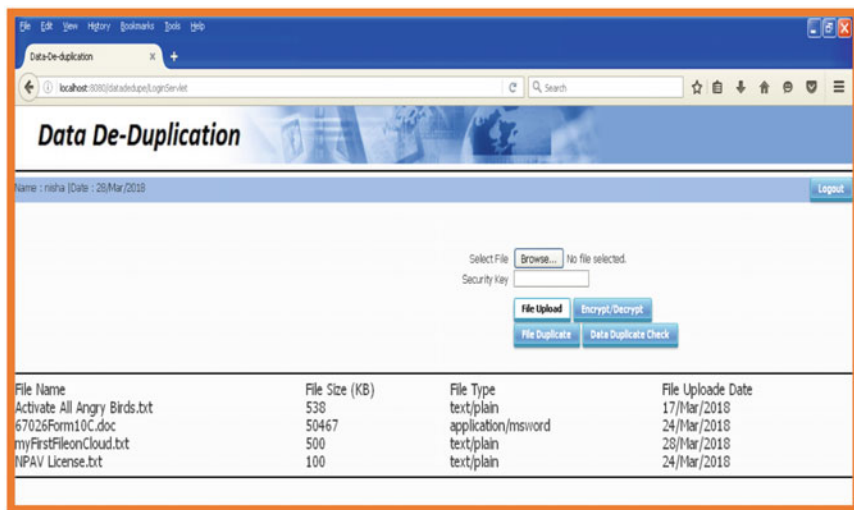


Fig. 9 File upload process

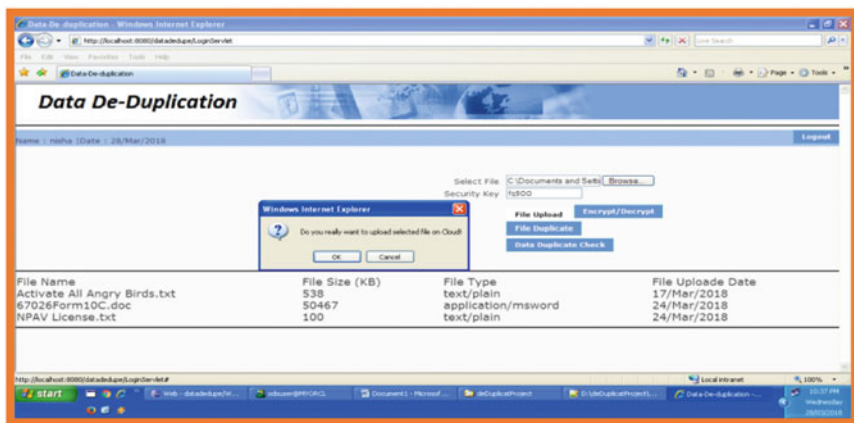


Fig. 10 File upload permission

(3) File uploads successfully as below.

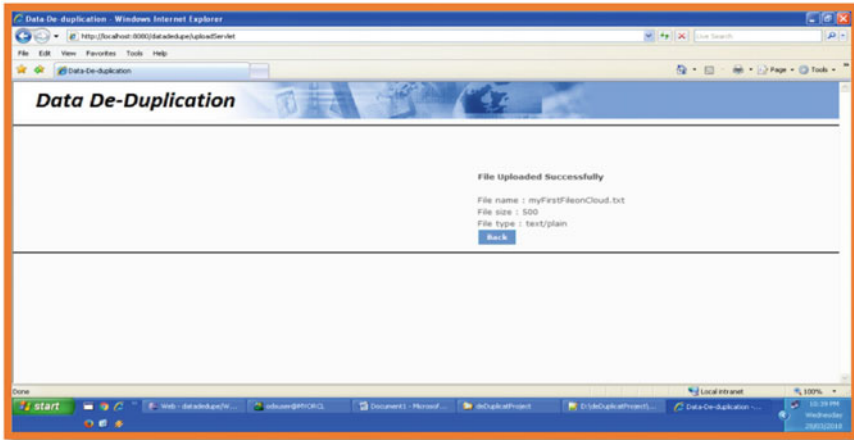


Fig. 11 File upload successfully showing size and type of file

This scenario in Fig. 11 shows File Uploaded Successfully in this output name of file, size of file and type of file which will upload on cloud.

(4) Uploaded file is getting display in the below given grid (Fig. 12).

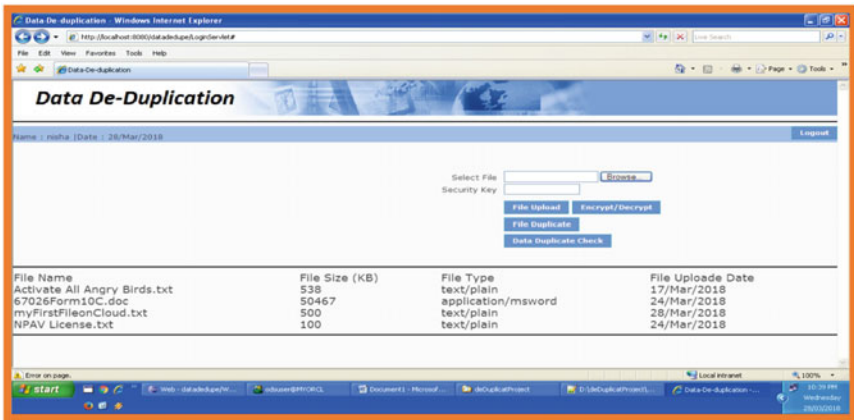


Fig. 12 File upload successfully display in list

In Fig. 13, uploaded file is getting display in the given grid. In this grid, we find that first column is file name, second column file size, third column file type, and last column file uploaded date on cloud.



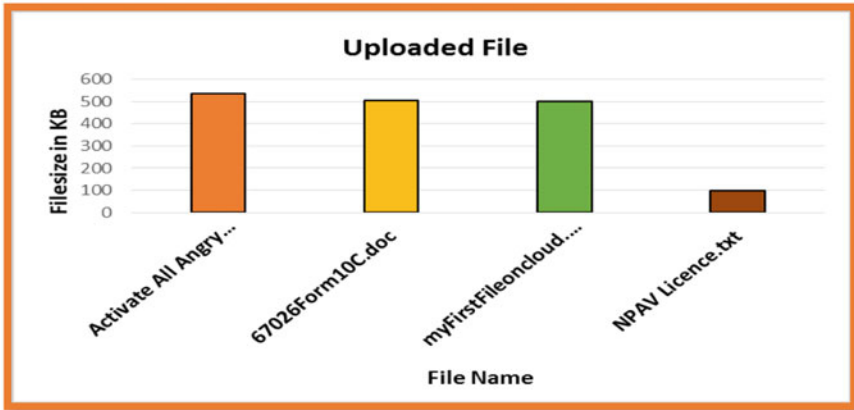


Fig. 13 Actual size of file

Figure 13 shows the actual size of file before the file/data deduplication process on the cloud. Uploaded file is getting display into the below given table (Fig. 14).

| DE_USER_NO | DE_FILE_SECURITY_KEY | DE_FILE_NAME | DE_FILE_SIZE | DE_FILE_TYPE | DE_FILE_UPLOAD_DATE | DE_FILE_DOWNLOAD_DATE | DE_FILE_ENCRYPT |
|------------|----------------------|------------------------------|--------------|--------------------|---------------------|-----------------------|-----------------|
| 1 | 4345f5g | Activate All Angry Birds.txt | 538 | text/plain | 17/Mar/2018 | 17/Mar/2018 | N |
| 2 | 45645fj | NPAY Licence.txt | 100 | text/plain | 24/Mar/2018 | 17/Mar/2018 | N |
| 3 | 67674gj | 67026Form10C.doc | 50467 | application/msword | 24/Mar/2018 | 17/Mar/2018 | N |

Fig. 14 Uploaded file in database

File is getting uploaded on below given cloud folder “upload” (Fig. 15).

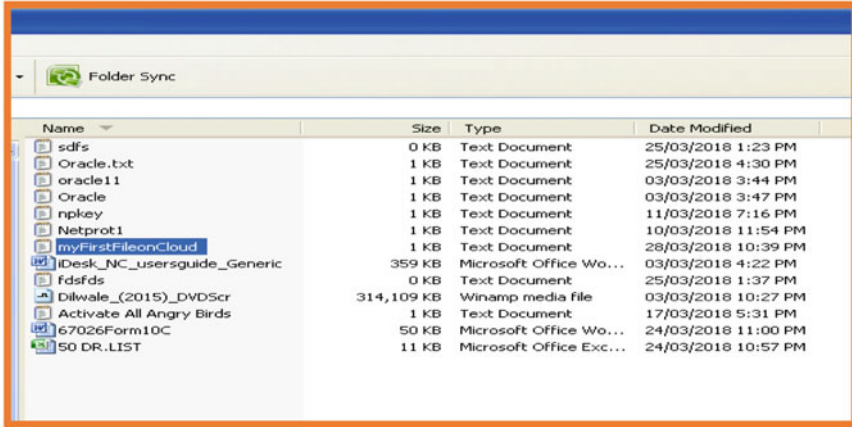


Fig. 15 File upload in upload folder

(1) File Upload Functionality

In Fig. 16, upload functionality, we check the file name and file duplicate check through check action Y/N.

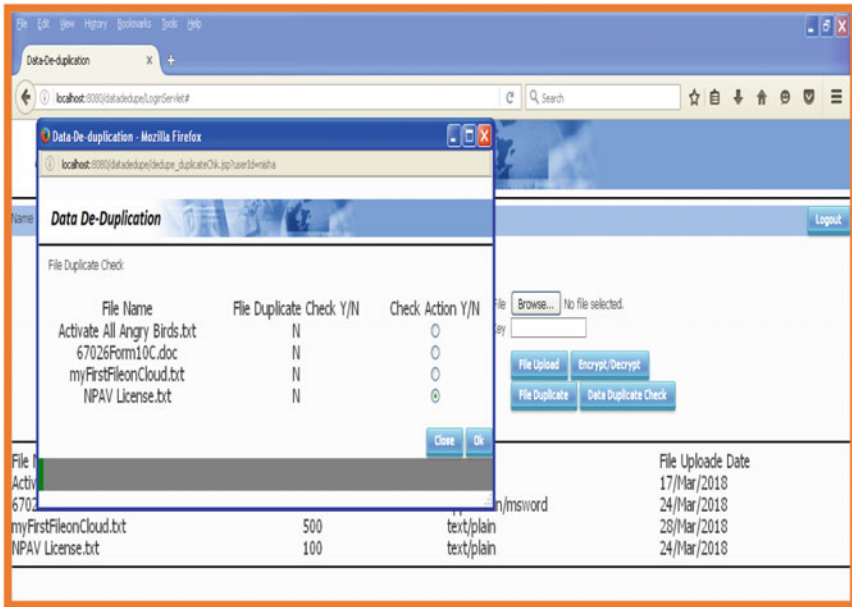


Fig. 16 File upload functionality



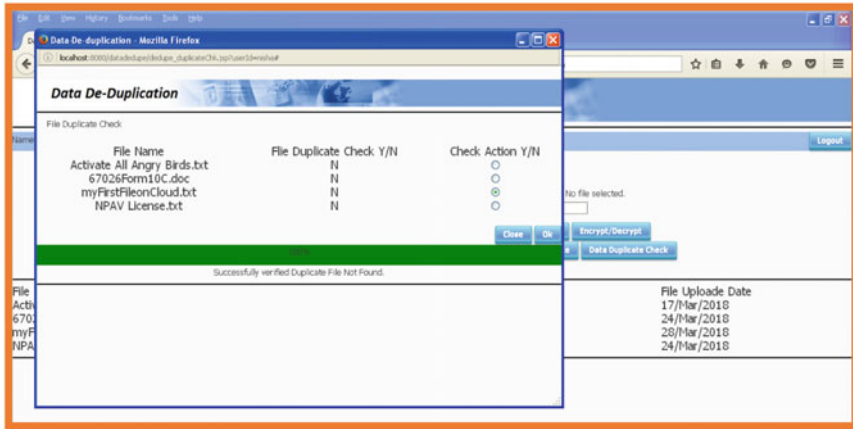


Fig. 17 File duplicate check

In Fig. 17, the outcome of the process of file duplicate check on cloud is shown along with progress bar that no duplicate file on cloud is found.

6 Conclusion

In this paper, we are applying data deduplication method which removes the duplicate file and duplicate data inside the file which helps improve the performance of the system. In the cloud, we provide encryption and decryption technique after removing the duplicate data or file which secure the user data. Security analysis defines the secure data from insider and outsider as well as data redundancy.

References

1. Pachpor, N. N., & Prasad, P. S. (2018). Improving the performance of system in cloud by using selective deduplication. In *IEEE 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*.
2. Mao, B., Jiang, H., Wu, S., & Tian, L. (2016). Leveraging data deduplication to improve the performance of primary storage systems in the cloud. *IEEE Transactions on Computers*, 65(6), 1775–1788.
3. Koller, R., & Rangaswami, R. (2010). I/O deduplication: Utilizing content similarity to improve I/O performance. In *Proceedings of USENIX File Storage Technologies*, February 2010 (pp. 1–14).
4. Meyer, D. T., & Bolosky, W. J. (2011). A study of practical deduplication. In *Proceedings of 9th USENIX Conference on File Storage Technologies*, February 2011 (pp. 1–14).



5. Clements, T., Ahmad, I., Vilayannur, M., & Li, J. (2009). Decentralized deduplication in SAN cluster file systems. In *Proceedings of USENIX Annual Technical Conference*, June 2009 (pp. 101–114).
6. Bibawe, C. B., & Baviskar, V. (2017). Secure authorized deduplication for data reduction with low overheads in hybrid cloud. *International Journal of Innovative Research in Computer and Communication Engineering*, 5(2), 1797–1804.
7. Jin, K., & Miller, E. L. (2009). The effectiveness of deduplication on virtual machine disk images. In *Proceedings of the Israeli Experimental Systems Conference*, May 2009 (pp. 1–12).
8. Srinivasan, K., Bisson, T., Goodson, G., & Voruganti, K. (2012). iDedup: Latency-aware, inline data deduplication for primary storage. In *Proceedings of 10th USENIX Conference on File Storage Technologies*, February 2012 (pp. 299–312).
9. Gode, R. V., & Dalvi, R. A survey on authorized deduplication technique for encrypted data with DARE scheme in a twin cloud environment. *IJIRCCE*, ISSN(Online): 2320-9801
10. El-Shimi, A., Kalach, R., Kumar, A., Oltean, A., Li, J., & Sengupta, S. (2012). Primary data deduplication-large scale study and system design. In *Proceedings of USENIX Annual Technical Conference*, June 2012 (pp. 285–296).
11. Kiswany, S., Ripeanu, M., Vazhkudai, S. S., & Gharaibeh, A. (2008). STDCHK: A checkpoint storage system for desktop gridcomputing. In *Proceedings of 28th International Conference on Distributed Computing Systems*, June 2008 (pp. 613–624).
12. Meister, D., Kaiser, J., Brinkmann, A., Cortes, T., Kuhn, M., & Kunkel, J. (2012) A study on data deduplication in HPC storage systems. In *Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis*, November 2012 (pp. 1–11).

Implementation of Collaborative Filtering for Product Recommendation in E-Commerce to Enhance Scalability and Performance



Niti Vishwas, Tapajyoti Deb, Ashim Saha and Lalita Kumari

Abstract With the advancement of technology and the Internet facilities, e-commerce business is growing very fast. E-commerce is trending across geographical boundaries. A user can place order sitting at home, and the product is delivered to the given address within the specified time period. At the same time, user is given a variety of options to choose from, thus making e-commerce a much convenient way of shopping. E-commerce applications nowadays have millions of users worldwide and billions of products to offer. Thus, it becomes difficult for a user to find a product of their choice from millions of available choices. Recommendation plays an important role in e-commerce applications. A recent study on the concerned topic has advocated for a recommendation system which can be a potential solution to the mentioned problem. This system uses various parameters such as users' purchase history, products in the cart, users' ratings and review to recommend an item to the target user. A good recommendation system is one which helps the user to find the appropriate product and also helps the organization to grow vertically as well as horizontally. Further, collaborative filtering is an algorithm which may be applied to product recommendation which successfully satisfies customer's needs and at the same time also helps in organization's growth. One of the biggest challenges these days is to provide efficiency and scalability while handling a large number of users and products. The research work studies and discusses different approaches of implementation of collaborative filtering for product recommendation system. The study also shows how to overcome the drawbacks of user-based collaborative filtering by implementing item-based collaborative filtering. This research work also demon-

N. Vishwas (✉) · T. Deb · A. Saha · L. Kumari
National Institute of Technology, Agartala, India
e-mail: niti.vishwas@gmail.com

T. Deb
e-mail: tapajyotideb@gmail.com

A. Saha
e-mail: ashim.cse@nita.ac.in

L. Kumari
e-mail: kumari12003@yahoo.co.in

strates how collaborative filtering can be implemented for big data using Hadoop, to overcome scalability problem and enhance performance.

Keywords E-commerce · Recommendation system · Collaborative filtering · Big data · Hadoop

1 Introduction

Recommendations play an important role in e-commerce applications. E-commerce commonly known as electronic commerce means business running through the Internet. E-commerce has been in existence since 1990 but getting popularity these days due to the advancement of the Internet facilities and successful e-commerce Web sites like Amazon, Flipkart, and Myntra [1]. E-commerce is trending across geographical boundaries. A user can place order sitting at home, and the product is delivered to the given address within the specified time period. At the same time, user is given a variety of options to choose from, thus making e-commerce a much convenient way of shopping. With the advancement of technology, e-commerce business is growing very fast leading to millions of users and billions of products. Hence, searching a product of one's choice is very difficult from a large number of products.

Recommender systems provide a solution to such a problem. Recommendation system provides customer satisfaction as well as helps in success of e-commerce business application. These systems use various parameters such as users' purchase history, products in the cart, users' ratings and review to recommend an item to target user. Collaborative filtering has been a successful recommendation algorithm since a long period of time. It is a similarity-based recommendation algorithm. There are two types of collaborative user-based filtering and collaborative filtering based on items.

There exist various challenges for a successful recommender system such as scalability, sparsity and accuracy. However, collaborative filtering helps to overcome all these challenges. One of the biggest challenges these days is to provide efficiency and scalability while handling a large number of users and products. This large volume of data which is beyond processing power and storage capacity of a system is referred to big data. Hadoop is a solution to big data. Collaborative filtering can also be implemented with Hadoop to overcome scalability as well as efficiency problems.

2 Literature Survey

Collaborative filtering technique is widely used in recommender systems. Collaborative filtering has been the most promising technique till date. There exist two types of collaborative filtering, item-based collaborative filtering and user-based collabora-

tive filtering [2]. Mulik and Gawali proposed importance of recommendation system to suggest item for the customer such as which movie to watch or what music to listen [3]. They implemented item-based collaborative filtering and found it effective and efficient.

Zhiyang Jia and Wei Gao proposed the recommender system as an online application which is capable of customizing the list of preferred attractions for the tourist [4]. Riyaz and Varghese proposed a scalable product recommendations using collaborative filtering in Hadoop for big data. They suggested an optimized HBase gives better performance. For low latency applications, HBase is highly preferred because of distributed architecture and it leverages the power of Apache Hadoop [5].

Xiaolong Xu, Lianyong Qi, Wanchun Dou, Xuyun Zhang, Chunhua Hu, Jiguo Yu, Yuming Zhou proposed structural balanced theory for e-commerce recommendation [6].

3 Different Approaches for Implementation of Collaborative Filtering in E-Commerce

3.1 *Item-Based Collaborative Filtering Recommendation Algorithms, 2001 by, George Karypis, John Riedl, Badrul Sarwar and Joseph Konstan*

The tremendous growth in a number of users and amount of products produces some challenges to recommender systems such as improvement in quality of recommendations, scalability and performance problem. Collaborative filtering is the most promising recommender system till date. It consists of collaborative filtering based on users and collaborative filtering based on items. This paper shows that collaborative filtering works better than collaborative filtering based on users. Collaborative filtering based on items works in two steps, computing the similarity of items and computing the prediction [7].

3.1.1 Item Similarity Computation

First step in item-based collaborative filtering is to determine similarity between items. Similarity between two items i and j can be estimated by determining the set of users who has rated item i and users who have rated item j and then applying similarity algorithm on it. Three similarity algorithms are listed below.

1. Cosign-based Similarity: In this algorithm, two items are considered as two vectors and similarity between items is determined by calculating cosign of angle between these vectors.

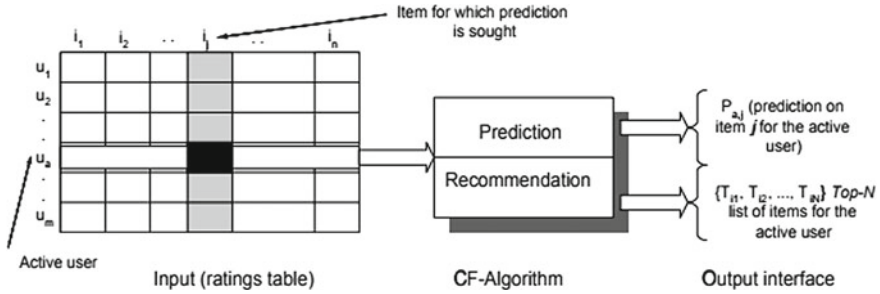


Fig. 1 Collaborative filtering process [7]

2. Correlation-based Similarity: This algorithm allows finding similarity between two items using Pearson correlation.
3. Adjusted Cosine Similarity: This algorithm overcomes the drawback of cosign similarity algorithm by considering the difference in rating scale between different users.

3.1.2 Prediction Computation

After generating a set of most similar items based on similarity calculations, the next step is to use target user ratings for prediction calculations to generate the most appropriate item for the user. For prediction calculations, the following techniques can be used.

1. Weighted Sum: This algorithm is used for predicting rating on target item I by calculating sum of ratings given by user on products similar to i .
2. Regression: This method makes use of regression model for approximation of ratings (Fig. 1).

Advantage(s):

1. Item-based collaborative filtering is more useful than user-based collaborative filtering.
2. Item-based collaborative filtering is static and gives better performance.
3. With model-based filtering, quality of recommendation system improves even with small datasets, hence taking less time for computation.

Disadvantage(s):

1. Item based collaborative filtering fails to generate recommendations when similarity between two items is null.

3.2 Recommendation System: Online Movie Store, 2013 by Archana T Mulik and S. Z. Gawali

Recommendation systems are needed for customer satisfaction as well as organizations' growth. Some of the popular recommendation systems consist of collaborative filtering, cluster model and search-based methods [3].

Collaborative filtering: Here, customer is represented as n-dimensional vector of items, where n is the number of items. Vector consists of two components, positively rated items and negatively rated items. Recommendation is generated using similarity of customer.

Cluster model: Customers are divided into different segments based on the similar trends among them. Then, the user is assigned to the segment containing most similar customer and the products purchased by the similar customers are recommended to the user.

Search-based methods: Search-based methods generate recommendation-based search for related items. The algorithm generates recommendations for the items based on the company or the keywords of the product which user has purchased and rated earlier.

Recommendation systems are based on rating prediction and ranking. However, item sequence generation technique and consumer-/manufacturer-oriented ranking techniques help to improve the quality of recommendation.

3.2.1 Algorithms for Rating Prediction

1. Item-based collaborative filtering technique: In this technique, similarity between the set of items that user has rated earlier and target item is calculated based on the results set of most similar items for recommendation.
2. Item similarity computation: In this technique, similarity between target item i and item j is calculated using the user average rating.
3. Prediction computation: Weighted sum approach is used in this technique to generate predictions.
4. Content-based technique: In this technique, recommendations are calculated based on users' profile, i.e., items that user has preferred in the past. Rating for user u on item s is assigned based on rating of user c on item s using cosine similarity measure.

3.2.2 Item Ranking

The unknown predicted ratings are used for ranking. Ranking is done on a scale of 1–5. Generally, items with ranking 3.5 and more are considered as highly ranked items. Items with higher ranks are more accurate for recommendation.

3.2.3 Item Sequence Generation Technique

User's next preference can be generated using Markov chain model which is based on last sequential data.

3.2.4 Consumer-/Manufacturer-Oriented Ranking Technique

This technique is used to filter reviews from a large number of reviews and find out the best suitable review.

Advantage(s):

1. Recommender systems provide customer satisfaction.
2. Item sequence generation technique helps to maintain recommendation accuracy.
3. Consumer-/manufacturer-oriented ranking technique helps to improve the quality of recommendation.

Disadvantage(s):

1. Scalability remains a problem when large dataset is used for processing.

3.3 *Item-Based Collaborative Filtering Approach for Big Data Application, 2014 by K. Sudha, M. Lavanya, A. Kanimozhi*

The problem in the existing system is that a large number of services are analyzed, leading to a high calculation time. A simple solution is to reduce the number of services to be processed in real time. Clustering is such a technique that similar services can reduce the size of the data. Collaborative filtering based on items is carried out with Hadoop to process large datasets. Item-based collaborative filtering is implemented with Hadoop to process large datasets. Clustering and item-based collaborative filtering are implemented as follows. The data needs to be uploaded to Hadoop system which clusters the data and then similarity algorithm is applied [8].

3.3.1 Clustering

Cluster models are one of the earliest recommender system techniques. Normally, users are clustered based on their browsing history and purchasing trends and products are recommended to similar users in the same cluster. Here, agglomerative

hierarchical clustering (AHC) algorithm is used for clustering. This algorithm uses of properties of item to cluster rather than user. AHC algorithm helps to achieve efficiency by reducing processing time.

3.3.2 Item-Based Collaborative Filtering

One of the model-based collaborative filtering is item-based collaborative filtering. It consists of two stages. First stage consists of calculation of similarity between items using similarity methods like correlation-based similarity or enhanced rating similarity. Second stage consists of prediction of ratings for unknown items.

3.3.3 Computing Similarity

It can be done using the following methods.

1. Pearson (correlation)-based similarity: This mechanism tends to measure correlation between two items and results in 1 if there exists a correlation and 0 otherwise.
2. Enhanced rating similarity: Similarity between item i and j is estimated by calculating similarity between users who have rated item i and users who have rated item j .

3.3.4 Select Neighbors

Neighbors of item i are calculated based on enhanced rating similarities between items using constraint formula.

3.3.5 Compute Predicted Rating

Based on the predicted rating, target item is recommended to the user. Rating gives better understanding of product being liked or not.

Advantage(s):

1. Clustering technique helps in reducing computation time for processing huge dataset.
2. Item-based collaborative filtering helps to achieve accuracy.

Disadvantage(s):

1. Item based collaborative filtering does not work when similarity algorithm results zero or null.

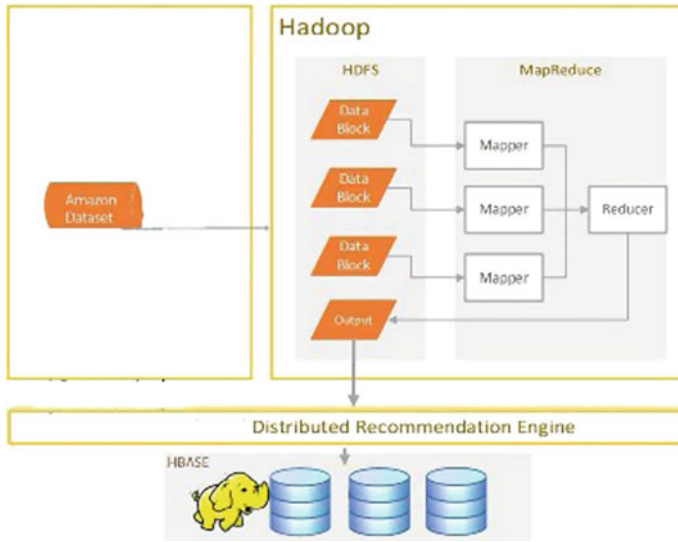


Fig. 2 System architecture [7]

3.4 A Scalable Product Recommendation Using Collaborative Filtering for Big Data Using Hadoop, 2015 by, Surekha Mariam Varghese and Riyaz PA

With the advancement of technology and the Internet facility, data is growing at an exponential rate nowadays. This huge amount of data needs a use of next-generation database. The amount of products and customers is growing rapidly leading to big data problems. Recommender systems play an important role in e-commerce. Traditional recommender systems often lack scalability and efficiency due to huge amount of data. Implementation of collaborative filtering using Apache Hadoop gives a solution to such problems [5] (Fig. 2).

Implementation of recommendation system consists of the following steps.

3.4.1 Data Extraction

Amazon product dataset is used as an input. Product dataset consists of reviews. Each review has a customer id and a rating associated with it. This dataset is loaded into Hadoop, and key-value pair is generated.

3.4.2 Data Analysis

This step aims to build a big data analysis system with recommendation system implemented on top of Hadoop. Python with MapReduce is used to build a scalable system.

Algorithm Realization

Realization of collaborative filtering algorithm consists of three steps:

1. Collect the user preferences: First, data is extracted from product dataset; then, data is grouped and preferences are made based on historical information. User preference is transformed into simple triple.
2. Find the similar items based on the user tastes: Pearson correlation coefficient (PCC) measure is used to calculate similarity between items.
3. Calculate the recommendations: Finally, the recommendations are calculated based on above two steps.

Data Storage

HBase is used as data storage. It is used to store recommendations. ZooKeeper is used to coordinate all activities in HBase. HBase helps to overcome latency problem in e-commerce applications.

Advantage(s):

1. Implementation of collaborative filtering with Hadoop helps to overcome scalability problem.
2. HBase used as storage provides low latency which is highly recommended.
3. Hadoop allows adding more data nodes as the size of data increases.

Disadvantage(s):

1. This algorithm does not work when target user does not share common user or common item with the database of e-commerce application.

3.5 Data Analytics Using Hadoop Framework for Effective Recommendation in E-Commerce Based on Social Network Knowledge, 2016, by Khyati P Raval and Dr. ShayamalTanna

Recommendations are very much important in e-commerce. Earlier recommender systems face many problems such as cold start and user preferences. Hence, collaborative filtering and clustering techniques have been developed to overcome such

problems. In this paper, recommender system has been implemented using social network data to generate more accurate recommendations [9].

The steps to implement the algorithm are as follows:

1. Start the process. Collect target users and their friends' information from social network Web site.
2. Collect purchase and rating history of target user from e-commerce Web site.
3. Find a set of closest friend using analytical hierarchy process (AHP) based on parameters such as place, age and language.
4. Select 100% or 50% of friends from the set of closest friends.
5. Implement hybrid collaborative filtering (memory-based + model-based) to generate recommendation list.
6. Finally, the recommendation list is shown to the user.

Advantage(s):

1. This recommendation system provides more accuracy.
2. This system provides scalability as well as efficiency since implemented with Hadoop

Disadvantage(s):

1. This recommendation system is not appropriate when user does not uses social media platform.

4 Conclusion

Recommendation system is very important in e-commerce applications. It helps the user to select a product of their choice ultimately leading to success of e-business. Collaborative filtering has been a successful recommendation system since a decade and is still promising by fulfilling all the needs. This paper shows different methods of implementing collaborative filtering and also shows how to overcome the drawback of user-based collaborative filtering by implementing item-based collaborative filtering.

Further structural balanced theory of product recommendation can be implemented in the future to overcome the failure of collaborative filtering in case of the absence of common user or common product, as due to sparsity of data there may be cases when both common user and common items are absent. Future work aims on implementing structural balanced theory of product recommendation which can overcome the problem by implementing enemy's enemy is a friend rule to find set of possible friends.

References

1. Importance of E-commerce. <http://webdesign.vinsign.com/what-is-ecommerce-importance.html>.
2. Yao, G., & Cai, L. (2016). *User-based and item-based collaborative filtering recommendation algorithms design*.
3. Mulik, A. T., & Gawali, S. Z. (2013). *Recommendation system: Online movie store*, Bharati Vidyapeeth College of Engineering: Pune.
4. Jia, Z., Gao, W., Yang, Y., & Chen, X. (2015). *User-based collaborative filtering for tourist attraction recommendations*.
5. Riyaz, P. A., & Varghese, S. M. (2015). *A scalable product recommendations using collaborative filtering in Hadoop for Bigdata*, M A College of Engineering.
6. Qi, L., Xu, X., Zhang, X., Dou, W., Hu, C., Zhou, Y., & Yu, J. (2015). *Structural balance theory-based E-commerce recommendation over big rating data*.
7. Karypis, G., Sarwar, S., Konstan, J., & Riedl, J. (2001). *Item-based collaborative filtering recommendation algorithms*, University of Minnesota.
8. Sudha, K. Lavanya, M., & Kanimozhi, A. (2014). *Itembased collaborative filtering approach for BigData application*, Kongunadu College of Engineering and Technology.
9. Raval, K. P., & Tanna, S. (2016). *Data analytics using Hadoop framework for effective recommendation in E-commerce based on social network knowledge*.

A Pre-emptive Goal Programming Model for Multi-site Production and Distribution Planning



Gaurav Kumar Badhotiya, Gunjan Soni and M. L. Mittal

Abstract The aim of every manufacturing organization is to fulfil the demand of their customers at minimum total cost or at maximum profit. To fulfil the demand of geographically dispersed customers, manufacturing organizations need to produce products at multiple manufacturing sites located close to the customer. Although planning function in such a scenario becomes very complex, this multi-site environment has been found to improve efficiency and provide better services to customers. It involves interlinked decisions involving procurement, production and distribution at different plants and distribution sites. Production and distribution planning are very important aspects of planning in multi-site manufacturing. To compete in this globalized market, enterprises should focus on optimizing and integrating production and distribution functions simultaneously. The authors in the research work present an integrated production and distribution planning problem for a two-level supply chain consisting of multiple manufacturing sites serving multiple selling locations. The problem is formulated as a multi-objective mixed-integer programming model considering three important aspects of production and distribution planning, viz. set-up, backorder and transportation capacity. Total cost, delivery time and backlog level are the three conflicting objectives that need to be minimized. The proposed multi-objective mathematical problem is solved using pre-emptive goal programming method. The performance of the proposed model is illustrated through an example problem instance. Further analysis is conducted to visualize the effect of changing priority level on objective function and deviation variable values.

Keywords Multi-site manufacturing · Production planning · Distribution planning · Goal programming

G. K. Badhotiya (✉) · G. Soni · M. L. Mittal
Department of Mechanical Engineering, Malaviya National Institute of Technology Jaipur, Jaipur,
Rajasthan, India
e-mail: gkb.choudhary@gmail.com

G. Soni
e-mail: gsoni.mech@mmit.ac.in

M. L. Mittal
e-mail: mlmittal.mnit@gmail.com

1 Introduction

The aim of every manufacturing organization is to fulfil the demand of their customers at minimum total cost or at maximum profit. To fulfil the demand of geographically dispersed customers, manufacturing organizations need to produce products at multiple manufacturing sites located close to customer. Although planning function in such a scenario becomes very complex, multi-site environment has been found to improve efficiency and provide better services to customers [1]. It involves interlinked decisions involving procurement, production and distribution at different plants and distribution sites [2]. For the introduction and structured literature review on production and distribution planning in multi-site manufacturing scenario, refer article [3].

Production and distribution planning are very important aspects of planning in multi-site manufacturing. It involves decisions related to determining production and inventory level at different facilities and quantity to be transported between facilities to fulfil the demand of customer with the objective of minimizing total cost or maximizing total profit. The focus of this study is on multi-site integrated production and distribution planning (MSIPDP) problem considering three important aspects: set-up, backorder and transportation capacity. It is observed from the literature that set-up cost for different products at manufacturing site and capacity of the transport vehicles are considered by [4] and [5], while backorder cost for unfulfilled demand and capacity of the transport vehicles are considered by [6] in their models. All the three important aspects of production and distribution planning in multi-site environment have not been considered in an integrated manner yet. Incorporating these three components represents the closeness to the real-life situation.

In the practical production and distribution planning decisions, there exist more than one conflicting objectives. Handling these multiple conflicting objectives simultaneously is a crucial task. Goal programming method developed by [7] and [8] is a well-known approach to solve multi-objective mathematical model. This approach looks for a solution which minimizes the deviation between achievement level and aspiration level of goals. A variance of goal programming method known as preemptive goal programming is implemented in this study.

The introduction, novelty and necessity of the problem considered are discussed in this section. Section 2 presents prior research works done in the selected area. Section 3 provides problem description and formulation. Computation results of the numerical example and analysis on the priority level of objectives are presented in Sects. 4, and 5 ends with the conclusion of the paper.

2 Literature Review

To handle multiple conflicting objectives, the goal programming method is implemented by several research studies in the literature. Reference [9] applied goal

programming method for coordinated production and logistics planning using FORTRAN computer programme. Reference [10] worked upon the coordination of production and marketing decisions and dealt with trade-off between conflicting objectives. Reference [11] studied production planning in a single chemical plant considering multiple conflicting goals of a practical environment.

There are several studies available that worked upon aggregate production planning (APP) problem using goal programming approach. A chance-constraint goal programming approach on APP with probabilistic demand pattern was taken into consideration by [12]. A polynomial goal programming approach on APP problem is applied by [13]. Reference [14] applied goal programming approach on multi-site APP problem for a lingerie company. The three multiple objectives considered are profit maximization, hiring and layoff cost minimization and import quota utilization. Reference [15] worked on aggregate production planning of perishable products with postponement strategy using pre-emptive goal programming method. Reference [16] considered resource utilization constraint in APP problem. The real data for the study was collected from a surface and materials science company.

An aggregate and detailed production planning problem in a multi-site manufacturing environment is considered by [17] and formulated as MIGP model with forward coverage policy of finished goods, inventory storage volume availability and overall cost minimization objectives. A multi-echelon supply chain logistic design and planning problem using weighted goal programming method is considered by [18] and [19].

3 Problem Description and Formulation

In this section, a general mixed-integer programming formulation is first presented and then converted into goal programming model. The problem involves production and distribution planning of multiple items i , multiple manufacturing sites s and multiple selling locations m over multiple time periods t as shown in Fig. 1. Fixed set-up cost and variable costs per unit of production at each manufacturing site; fixed vehicle cost and variable per unit transportation costs; backordering and inventory holding cost at facilities in the network have been considered in the model.

The following parameters, decision and auxiliary variables are used to formulate the multi-objective mixed-integer programming (MOMIP) and goal programming model.

Parameters

| | |
|-------------|--|
| $D_i^{m,t}$ | Demand of product i at selling location m during period t |
| St_i^s | Set-up time of product i at manufacturing site s |
| Pt_i^s | Processing time of product i at manufacturing site s |
| Pc_i^s | Production cost per unit of product i in manufacturing site s |
| Hs_i^s | Inventory holding cost product i at manufacturing site s per time period |

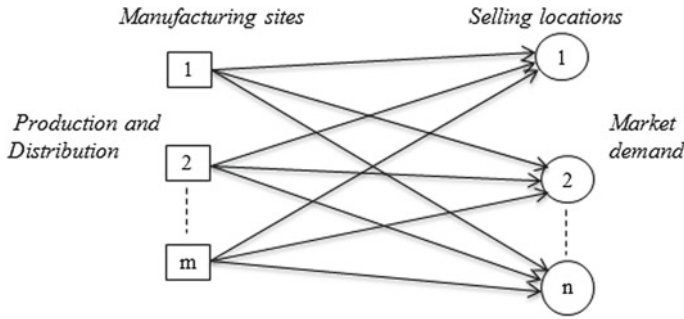


Fig. 1 Overview of multi-site production and distribution network

- Hm_i^m Inventory holding cost of product i at selling location m per time period
- $Vt_i^{s,m}$ Variable transportation cost of product i from manufacturing site s to selling location m
- $Ft^{s,m}$ Fixed transportation cost between manufacturing site s and selling location m
- Sc_i^s Set-up cost of product i at manufacturing site s
- Bc_i^m Backordering cost of product i at selling location m
- $Pcp^{s,t}$ Production capacity of manufacturing site s in period t
- Ss_i^s Maximum storage capacity of manufacturing site s for product i
- Sm_i^m Maximum storage capacity of selling location m for product i
- $TC^{s,m}$ Maximum capacity of transport vehicle
- M A sufficient large number
- θ Fraction of demand that is allowed to be backordered
- $DT^{s,m}$ Transportation time from manufacturing site s to selling location m

Decision Variables

- $x_i^{s,t}$ Quantity of product i produced at manufacturing site s during period t
- $Qt_i^{s,m,t}$ Quantity of product i transported from manufacturing site s to selling location m in time period t
- $Qb_i^{m,t}$ Backorder of product i at selling location m in time period t
- $Qs_i^{s,t}$ Inventory level of product i at manufacturing site s at the end of the period t
- $Qm_i^{m,t}$ Inventory level of product i at selling location m at the end of time period t
- $bs_i^{s,t}$ Binary variable denoting set-up of manufacturing site s for producing product i in period t
- $bt^{s,m,t}$ Binary variable denoting transportation between manufacturing site s and selling location m in period t
- Al_1 Aspiration level of total cost goal
- Al_2 Aspiration level of delivery time goal
- Al_3 Aspiration level of backorder level goal

Auxiliary Variables

- Al_1^+ Deviation of overachievement of Al_1
 Al_1^- Deviation of underachievement of Al_1
 Al_2^+ Deviation of overachievement of Al_2
 Al_2^- Deviation of underachievement of Al_2
 Al_3^+ Deviation of overachievement of Al_3
 Al_3^- Deviation of underachievement of Al_1

The objective function of the MOMIP model is:

$$\begin{aligned}
 \text{Min. } Z_1 = & \sum_i \sum_s \sum_t Pc_i^s x_i^{s,t} + \sum_i \sum_s \sum_t Hs_i^s Qs_i^{s,t} + \sum_i \sum_m \sum_t Hm_i^m Qm_i^{m,t} \\
 & + \sum_s \sum_m \sum_t Ft^{s,m} bt^{s,m,t} + \sum_i \sum_s \sum_m \sum_t Vt_i^{s,m} Qt_i^{s,m,t} \\
 & + \sum_i \sum_s \sum_t Sc_i^s bs_i^{s,t} + \sum_i \sum_m \sum_t Bc_i^m Qb_i^{m,t} \quad (1)
 \end{aligned}$$

$$\text{Min. } Z_2 = \sum_i \sum_s \sum_m \sum_t (Dt_s^{m,t} / TC^{s,m}) Qt_i^{s,m,t} \quad (2)$$

$$\text{Min. } Z_3 = \sum_i \sum_m \sum_t Qb_i^{m,t} \quad (3)$$

The constraints of the proposed MOMIP model are:

$$Qs_i^{s,t} = Qs_i^{s,t-1} + x_i^{s,t} - \sum_m Qt_i^{s,m,t} \quad \forall i, s, t \quad (4)$$

$$(Qm_i^{m,t-1} - Qb_i^{m,t-1}) - (Qm_i^{m,t} - Qb_i^{m,t}) = D_i^{m,t} - \sum_s Qt_i^{s,m,t} \quad \forall i, m, t \quad (5)$$

$$x_i^{s,t} \leq Mbs_i^{s,t} \quad \forall i, s, t \quad (6)$$

$$x_i^{s,t} Ft^{s,m} + St_i^s bs_i^{s,t} \leq Pcp^{s,t} \quad \forall i, s, t \quad (7)$$

$$Qt_i^{s,m,t} \leq TC^{s,m} bt^{s,m,t} \quad \forall i, s, m, t \quad (8)$$

$$Qb_i^{m,t} \leq \theta D_i^{m,t} \quad \forall i, m, t \quad (9)$$

$$Qs_i^{s,t} \leq Ss_i^s \quad \forall i, s, t \quad (10)$$

$$Qm_i^{m,t} \leq Sm_i^m \quad \forall i, m, t \quad (11)$$

$$Qs_i^{s,0}, Qm_i^{m,0}, Qb_i^{m,0}, Qb_i^{m,T} = 0 \quad (12)$$

$$bs_i^{s,t}, bt^{s,m,t} \in \{1, 0\} \quad (13)$$

$$x_i^{s,t}, Qs_i^{s,t}, Qm_i^{m,t}, Qt_i^{s,m,t}, Qb_i^{m,t} \geq 0 \text{ and integer} \quad (14)$$

The objective functions (i) minimizing the total cost comprising production cost and inventory holding cost at different production site and selling locations, fixed and variable transportation cost between plants and selling locations, set-up cost at manufacturing sites and backorder cost, (ii) minimizing the total delivery time between manufacturing sites and selling locations and (iii) minimizing the total backorder level representing the total unfulfilled demand of selling locations are represented by Eqs. (1)–(3). Constraints (4) and (5) are inventory balance equation at each manufacturing site and selling locations, respectively. Constraint (6) ensures that set-up cost is incurred only when production is happening at a site. Constraint (7) represents production capacity in terms of time at each manufacturing site. Constraint (8) ensures that transported quantity should be less than or equal to transportation capacity. Backorder quantity should be equal to or less than some fraction of demand represented by Constraint (9). Constraints (10) and (11) are storage capacity constraints at manufacturing sites and selling locations, respectively. Initial inventory at site and selling location and backorder quantity at start and end of period are assumed as zero, represented by constraint (12). Constraints (13) and (14) define the nature of the variables.

The above MSIPDP problem is transformed into mixed-integer pre-emptive goal programming model formulation as follows:

$$\text{Min } P_1 Al_1^+ + P_2 Al_2^+ + P_3 Al_3^+$$

Subjected to,

$$\begin{aligned} & \sum_i \sum_s \sum_t Pc_i^s x_i^{s,t} + \sum_i \sum_s \sum_t Hs_i^s Qs_i^{s,t} + \sum_i \sum_m \sum_t Hm_i^m Qm_i^{m,t} \\ & + \sum_s \sum_m \sum_t Ft^{s,m} bt^{s,m,t} + \sum_i \sum_s \sum_m \sum_t Vt_i^{s,m} Qt_i^{s,m,t} \\ & + \sum_i \sum_s \sum_t Sc_i^s bs_i^{s,t} + \sum_i \sum_m \sum_t Bc_i^m Qb_i^{m,t} - Al_1^+ + Al_1^- = Al_1 \end{aligned} \quad (15)$$

$$\sum_i \sum_s \sum_m \sum_t (Dt_s^{m,t} / TC^{s,m}) Qt_i^{s,m,t} - Al_2^+ + Al_2^- = Al_2 \quad (16)$$

$$\sum_i \sum_m \sum_t Qb_i^{m,t} - Al_3^+ + Al_3^- = Al_3 \quad (17)$$

Equations (4)–(14)

$$Al_1^+, Al_1^-, Al_2^+, Al_2^-, Al_3^+, Al_3^- \geq 0 \text{ and integer} \quad (18)$$

P_1, P_2 and P_3 ($P_1 > P_2 > P_3$) represent the level of priority for goals specified by management with P_1 as the highest priority and P_3 as the lowest. In this formulation,

all the objective functions' direction is minimization; therefore, the undesirable deviation variable in all objective functions is overachievement of the goal or positive deviation variable. The proposed MIGP model can be solved by using any linear programming solver.

4 Computational Results and Discussion

This section describes the problem instance taken and reports the analytical results provided by MIP solver. Due to the delicacy of the information for any industry, the data was collected through the literature and/or generated randomly. The proposed mathematical model was written and solved using CPLEX solver provided via IBM ILOG CPLEX 12.7 on a PC Intel Core i5 1.7 GHz and 4 GB RAM. The branch-and-cut algorithm of CPLEX terminates only if an optimal solution of the problem is found.

A problem instance is solved for 4 planning periods in which 3 products are produced in 3 manufacturing sites and transported to 4 selling locations. Each of the products can be stored at manufacturing sites and at selling locations according to the maximum storage capacity. Different parameters values as input data are shown in Table 1.

The target value or aspiration levels of objectives are computed by solving single-objective models. The computed target values are {784478.8; 91141; 0}. The mixed-integer goal programming model consists of 981 constraints and 531 variables out of which 180 are binary, 345 are integer, and 6 are deviation variables. The highest

Table 1 Input parameter values

| Parameter | Value |
|------------------------------|-----------------------------------|
| Demand | Uniform (120–200) |
| Set-up time | Uniform (500–1000) min/set-up |
| Processing time | Uniform (6–10) min |
| Production cost | Uniform (15–25) |
| Holding cost at site | Uniform (3–7) |
| Holding cost at market | Uniform (3–7) |
| Variable transportation cost | Uniform (1–3) |
| Fixed transportation cost | 500 |
| Set-up cost | 4500 |
| Backorder cost | Uniform (5–13) |
| Production capacity | Uniform (3000–4000) min |
| Transportation capacity | 100 |
| Delivery time | Uniform (6–32) hours per 100 unit |

Table 2 Quantity of production at manufacturing site

| Product | Manufacturing site | Time period | | | |
|---------|--------------------|-------------|-----|-----|-----|
| | | 1 | 2 | 3 | 4 |
| 1 | 1 | 300 | 330 | 0 | 279 |
| | 2 | 0 | 293 | 300 | 300 |
| | 3 | 0 | 0 | 0 | 0 |
| 2 | 1 | 481 | 492 | 0 | 0 |
| | 2 | 0 | 0 | 0 | 0 |
| | 3 | 0 | 438 | 426 | 0 |
| 3 | 1 | 300 | 300 | 275 | 300 |
| | 2 | 0 | 0 | 0 | 0 |
| | 3 | 0 | 319 | 312 | 0 |

Table 3 Quantity of backorder at selling location

| Product | Selling location | Time period | | | |
|---------|------------------|-------------|----|----|---|
| | | 1 | 2 | 3 | 4 |
| 1 | 1 | 31 | 12 | 55 | 0 |
| | 2 | 63 | 20 | 49 | 0 |
| | 3 | 58 | 21 | 42 | 0 |
| | 4 | 0 | 0 | 0 | 0 |
| 2 | 1 | 69 | 39 | 0 | 0 |
| | 2 | 38 | 0 | 0 | 0 |
| | 3 | 66 | 32 | 0 | 0 |
| | 4 | 0 | 0 | 0 | 0 |
| 3 | 1 | 80 | 28 | 0 | 0 |
| | 2 | 75 | 13 | 0 | 0 |
| | 3 | 62 | 0 | 0 | 0 |
| | 4 | 0 | 0 | 0 | 0 |

priority goal is minimization of overachievement of total cost followed by delivery time and backorder level. Computational results of decision variable values are shown in Tables 2, 3 and 4. Table 2 shows the production quantity at manufacturing site.

The quantities produced are transported to selling location according to the transportation capacity limit. When demand exceeds the inventory at selling location and transported quantity, then there comes a backorder. Table 3 shows the backorder quantity of each product at each selling location in each time period. Table 4 shows the quantities transported from manufacturing site to selling location. It can be seen that maximum transported quantity in a period is less than or equal to the capacity limit.

The results indicate that there is no deviation in total cost objective function. It shows that the optimum value of the goal is equal to the targeted value. The

Table 4 Quantity transported from manufacturing site to selling location

| Product | Manufacturing site | Selling location | Time period | | | |
|---------|--------------------|------------------|-------------|-----|-----|-----|
| | | | 1 | 2 | 3 | 4 |
| 1 | 1 | 1 | 100 | 100 | 0 | 79 |
| | | 2 | 100 | 96 | 0 | 100 |
| | | 3 | 100 | 100 | 34 | 100 |
| | | 4 | 0 | 0 | 0 | 0 |
| | 2 | 1 | 0 | 99 | 100 | 100 |
| | | 2 | 0 | 94 | 100 | 100 |
| | | 3 | 0 | 100 | 100 | 100 |
| | | 4 | 0 | 0 | 0 | 0 |
| | 3 | 1 | 0 | 0 | 0 | 0 |
| | | 2 | 0 | 0 | 0 | 0 |
| | | 3 | 0 | 0 | 0 | 0 |
| | | 4 | 0 | 0 | 0 | 0 |
| 2 | 1 | 1 | 100 | 100 | 100 | 28 |
| | | 2 | 100 | 100 | 84 | 0 |
| | | 3 | 100 | 100 | 100 | 61 |
| | | 4 | 0 | 0 | 0 | 0 |
| | 2 | 1 | 0 | 0 | 0 | 0 |
| | | 2 | 0 | 0 | 0 | 0 |
| | | 3 | 0 | 0 | 0 | 0 |
| | | 4 | 0 | 0 | 0 | 0 |
| | 3 | 1 | 0 | 100 | 100 | 100 |
| | | 2 | 0 | 100 | 100 | 64 |
| | | 3 | 0 | 100 | 100 | 100 |
| | | 4 | 0 | 0 | 0 | 0 |
| 3 | 1 | 1 | 100 | 100 | 100 | 100 |
| | | 2 | 100 | 100 | 100 | 100 |
| | | 3 | 100 | 100 | 75 | 100 |
| | | 4 | 0 | 0 | 0 | 0 |
| | 2 | 1 | 0 | 0 | 0 | 0 |
| | | 2 | 0 | 0 | 0 | 0 |
| | | 3 | 0 | 0 | 0 | 0 |
| | | 4 | 0 | 0 | 0 | 0 |
| | 3 | 1 | 0 | 100 | 100 | 13 |
| | | 2 | 0 | 100 | 100 | 27 |
| | | 3 | 0 | 100 | 91 | 0 |
| | | 4 | 0 | 0 | 0 | 0 |

Table 5 Effect of priority level on objective function values

| Run | Objective function (priority) | | | Deviation | | | | | |
|-----|-------------------------------|---------------------|-----------------|-----------|----------|----------|----------|----------|----------|
| | Total cost | Total delivery time | Backorder level | AI_1^+ | AI_1^- | AI_2^+ | AI_2^- | AI_3^+ | AI_3^- |
| 1 | 784478.8 (P1) | 104414 (P2) | 853 (P3) | 0 | 0 | 13273 | 0 | 853 | 0 |
| 2 | 784478.8 (P1) | 104414 (P3) | 853 (P2) | 0 | 0 | 13273 | 0 | 853 | 0 |
| 3 | 808925.8 (P2) | 91173 (P1) | 522(P3) | 24447 | 0 | 32 | 0 | 522 | 0 |
| 4 | 786519.4 (P2) | 109639 (P3) | 214 (P1) | 2040.6 | 0 | 18498 | 0 | 214 | 0 |
| 5 | 842230.7 (P3) | 91141 (P1) | 26 (P2) | 57751.9 | 0 | 0 | 0 | 26 | 0 |
| 6 | 845221 (P3) | 91141 (P2) | 0 (P1) | 60742.2 | 0 | 0 | 0 | 0 | 0 |

second-priority goal has positive deviation of 13,273, and the third-priority goal also has positive deviation of 853, which indicates over satisfied value of goal. Further analysis to find out the effect of priority level on objective function values is also conducted, and the results are demonstrated in Table 5.

The feasible solutions obtained at different priority settings show the difference in objective function and deviation values. It can be seen from Table 5 that in the first two scenarios minimizing total cost objective is the highest priority and the results obtained are having no deviation from the target values. The same can be observed for other objective values on different priority settings. These values show the difference of solution from the target values generally decided by the management team. After analysing the results, it can be stated that focusing or providing high priority to one objective may not necessarily cause an improvement in another objective value even if they are having same direction.

5 Conclusion

In this paper, the production and distribution planning problem in a multi-site manufacturing environment is discussed. A multi-objective pre-emptive goal programming model for multi-product, multi-site scenario is presented to simultaneously optimize three objectives. Three conflicting objectives are minimization of total cost, total delivery time and backorder level. The computational results indicate the performance of the model and provide an estimate to management for different priority levels of objective functions. This paper provides a quantitative tool for management or decision-maker to analyse trade-off and priority consideration between multiple conflicting objectives. Further research work can be directed towards considering uncertainty in the parameters and objective functions.

References

1. Guinet, A. (2001). Multi-site planning: A transshipment problem. *International Journal of Production Economics*, 74(1), 21–32.
2. Darvish, M., Larrain, H., & Coelho, L. C. (2016). A dynamic multi-plant lot-sizing and distribution problem. *International Journal of Production Economics*, 54(22), 6707–6717.
3. Badhotiya, G. K., Soni, G., Mittal, M. L.: An analysis of mathematical models for multi-site production and distribution planning. *International Journal of Intelligent Enterprise* (2018) (Article in press).
4. Park, Y. B. (2005). An integrated approach for production and distribution planning in supply chain management. *International Journal of Production Research*, 43(6), 1205–1224.
5. Melo, R. A., & Wolsey, L. A. (2012). MIP formulations and heuristics for two-level production-transportation problems. *Computers and Operations Research*, 39(11), 2776–2786.
6. Entezaminia, A., Heydari, M., & Rahmani, D. (2016). A multi-objective model for multi-product multi-site aggregate production planning in a green supply chain: Considering collection and recycling centers. *Journal of Manufacturing Systems*, 40, 63–75.

7. Charnes, A., Cooper, W. W., & Ferguson, R. O. (1955). Optimal estimation of executive compensation by linear programming. *Management Science*, 1(2), 138–151.
8. Charnes, A., & Cooper, W. W. (1961). *Management models and industrial applications of linear programming*. New York: Wiley.
9. Lawrence, K. D., & Burbridge, J. J. (1976). A multiple goal linear programming model for coordinated production and logistics planning. *International Journal of Production Research*, 14(2), 215–222.
10. Taylor, B. W., & Anderson, P. F. (1979). Goal programming approach to marketing/production planning. *Industrial Marketing Management*, 8(2), 136–144.
11. Zanakis, S. H., & Smith, J. J. S. (1980). Chemical production planning via goal programming. *International Journal of Production Research*, 18(6), 687–697.
12. Rakes, T. R., Franz, L. S., & James Wynne, A. (1984). Aggregate production planning using chance-constrained goal programming. *International Journal of Production Research*, 22(4), 673–684.
13. Deckro, R. F., & Hebert, J. E. (1984). Goal programming approaches to solving linear decision rule based aggregate production planning models. *IIE Transactions*, 16(4), 308–315.
14. Leung, S. C., Wu, Y., & Lai, K. K. (2003). Multi-site aggregate production planning with multiple objectives: A goal programming approach. *Prod. Plan. Control.*, 14(5), 425–436.
15. Leung, S. C., & Ng, W. L. (2007). A goal programming model for production planning of perishable products with postponement. *Computer and Industrial Engineering*, 53(3), 531–541.
16. Leung, S. C., & Chan, S. S. (2009). A goal programming model for aggregate production planning with resource utilization constraint. *Computer and Industrial Engineering*, 56(3), 1053–1064.
17. Kanyalkar, A. P., & Adil, G. K. (2007). Aggregate and detailed production planning integrating procurement and distribution plans in a multi-site environment. *International Journal of Production Research*, 45(22), 5329–5353.
18. Khalili-Damghani, K., & Tajik-Khaveh, M. (2015). Solving a multi-objective multi-echelon supply chain logistic design and planning problem by a goal programming approach. *International Journal of Management Science and Engineering Management*, 10(4), 242–252.
19. Hafezalkotob, A., Khalili-Damghani, K., & Ghashami, S. S. (2016). A three-echelon multi-objective multi-period multi-product supply chain network design problem: A goal programming approach. *Journal of Optimization in Industrial Engineering*, 10(21), 67–78.

An Analysis of Comorbidities' Role in Diabetes Mellitus and Its Data-Intensive Technology-Based Prediction to Reduce Risk and Diagnostic Costs



M. Venkatesh Saravanakumar and M. Sabibullah

Abstract Present disease management system focuses on identifying single disease and its curable methods. This approach is not suitable for patients with diabetes in present and in future. Diabetes must be concerned with healthcare providers (health care constituents) with new and evolving attributes of diabetes risk factors (like gut micro-biota, Hct, Plt, Hgb, and MPV) through data-intensive technology (Big Data), which has capability to ensure the accuracy delivery in the line of diabetes patients' care prediction. Having an understanding of diabetes-related *comorbid conditions* is crucial when dealing with diabetes because *comorbidities* are known to significantly increase the risk of getting serious, and ultimately, the cost for the medication will increase. Comorbidity is the incidence of additional persistent conditions in the same patient with a prominent disease and occurs frequently among patients with diabetes. The main purpose of study is to identify the impact of *data-driven comorbidity effects in diabetes patients* by predicting their risk status through *data-intensive technology* (Big Data), as uncovered problem domain in computer application. Hence, this effort would open up more impacts to enhance the research potentials on the killer disease, diabetes mellitus.

Keywords Diabetes mellitus · Comorbidities · Data-intensive technology · Big data · Hadoop · Hive and R · Ecosystem · Health care

M. Venkatesh Saravanakumar (✉)

PG and Research Department of Computer Science, Sudharsan College of Arts and Science, Pudukkottai, TN, India

e-mail: Venkatesh.srivi@gmail.com

M. Sabibullah

PG and Research Department of Computer Science, Jamal Mohamed College (Autonomous), Tiruchirappalli, TN, India

e-mail: manavaisafi@yahoo.com

© Springer Nature Singapore Pte Ltd. 2020

M. Pant et al. (eds.), *Performance Management of Integrated Systems and its Applications in Software Engineering*, Asset Analytics,

https://doi.org/10.1007/978-981-13-8253-6_8

1 Introduction

1.1 Motivation of Research

Understanding the consequence of comorbidity indices' presence, its kind, and specific degree of medical healthcare exploitation is very much essential to get insights into prospective healthcare needs of diabetes patients. As per the growing number of diabetic patients and the complexities involved with the disease itself, in future, focusing on curing single disease is not suitable for patients with diabetes.

Present disease care program especially for diabetes must be added to provide better care modules which focus on multiple chronic diseases (like comorbidities). The core diabetes care modules should focus on diabetes-related comorbidities like cardiovascular diseases, retinopathy, nephropathy, and diabetic foot because patients with diabetes not only have diabetes but also they may have diabetes-related comorbidity and non-diabetes-related comorbidities, such as depression and musculoskeletal diseases.

To predict patients' risks while taking into consideration all comorbidities is a new way of medication. Changes can be made in healthcare cost structure, improving outcomes, and interventions to target high risks which must be considered and handled effectively. The main focus of study is to scrutinize the impact of *data-driven comorbid* indices in which the patients with diabetes will be analyzed by predicting their risk status through *data-intensive technology* (Big Data) to find mortality rates.

2 Diabetes Mellitus (DM) and Prevalence of DM

Diabetes mellitus (DM) is one of the common non-communicable diseases (NCDs) globally. It is the seventh or eighth leading disease which causes death even in developed countries [1, 2]. According to International Diabetes Federation (IDF), the number of people living with diabetes is expected to rise very high by 2035. IDF has also estimated that about 65.1 million people in India are living with diabetes. Diabetes can be classified into type 1, type 2, and gestational diabetes mellitus of which type 2 is the most common form of diabetes prevalent among the reported diabetes cases.

Chronic hyperglycemia is the key factor which causes the mortality and morbidity in type 2 diabetic patients [3]. The effect of chronic hyperglycemia leads to other complications which may be long term or short term. Due to poor knowledge and healthcare system, the actual number of persons affected by the diabetes is still unknown and the count may increase rapidly. The same issue is prevalent in India especially in village due to poor healthcare monitoring System.

Diabetes epidemic poses war threatening to the healthcare system in India. This global pandemic disease must be handled carefully and wisely. Diabetes is a chronic disorder which is evolved by high blood (glucose) sugar level. Accumulation of level

of blood sugar must be leveled and controlled by the body. If it is not done naturally by the body, it can cause serious health-related complications. The population of diabetes patients all over the world is doubled in the last three decades as it was unexpected. Due to Westernized lifestyle in India, this will go double in 2030 [1]. Diabetes has different pathophysiology and inference from individuals. The exact core mechanisms and their relationship which may lead to show symptoms are still unclear. Recently, diabetes is related to comorbid diseases such as vascular complications [2], kidney and its related functions' failures [2], dysfunctions of peripheral nerves [2], issues related to functionality of heart and the diseases [3, 4], cognitive behavior disorders [5, 6], disorder in the eyes [7], and increased level of pressure in blood [8]. Each of these comorbidities opens up a unique direction of research. Investigating the diabetic comorbidity indices will allow us to assess the strength of the relations among the different kinds of comorbidities. By doing so, it is easy to rank them according to their significance.

To make use of the full potential of "Big Data" for health care and in medical diagnosis, an effective methodology must be formulated to pull out clinically relevant features from large datasets of electronic health records (EHRs). In broader sense, analyses of comorbidities using EHR data of DM to be examined are essential in order to develop a new quantitative skeleton to measure all possible comorbidity relations for DM1 and DM2.

3 Literature Review

The present analysis can be found only in the limited works carried out by the authors using Hadoop and Hive and R technologies toward diabetes research. It was found that there are very few and limited papers on Big Data and diabetes research, whereas many authors proposed and shown experimental results for diabetes works carried out by using different machine learning algorithms. More works may be done and proposed by focusing on the uncovered areas related to diabetes and Big Data technology. The paper outlines the more thrust areas on DM research and potentials of Big Data and its tools, by which exploration of results would be possible.

Sharmila and Vethamanickam [9] presented an overview of various DM techniques and application of diabetic dataset using Hadoop platform. Saravana kumar et al. [10] proposed a methodology for diabetic data analysis with Hadoop ecosystem, which enables affordable healthcare. Srivastava et al. [11] explained diabetic prediction methods by using Hive and R. It was analyzed with different attributes and also employed KNN algorithm to have better accuracy. Sadhana and Shetty [12] proposed a model to analyze diabetic dataset using Hive and R.

Kavakiotis et al. [13] normalized machine learning methods with data mining algorithm and reviewed the machine learning algorithm in the field of diabetes research. Here, wide ranges of ML algorithms were employed. Alic et al. [14] adopted machine learning techniques to classify the diabetes and cardiovascular diseases. Sabibullah

[15] experimented a prognostic neural model for diabetic risks prediction using ANN classifier algorithm.

Farran et al. [16] designed a model to assess the risk of T2D with machine learning algorithm by taking the comorbidity and hypertension as measure of scale. Machine learning algorithms, namely logistic regression, KNN, multifactor dimensionality reduction, and SVM, were used.

Veena and Anjali [17] reviewed decision support system for predicting diabetes mellitus, and also reviewed different pre-processing techniques based on SVM. Srikanth and Deverapalli [18] studied the classification algorithm to diagnose diabetes. Classification algorithms like decision tree, Bayes algorithm, and rule-based algorithm have been analyzed and suggested that it can be further enhanced by using linear regression model or logistic regression model. Patil et al. [19] explored the association rule for classifying T1D patients and implemented in Apriori algorithm. Aishwarya and Anto [20] proposed a decision support system for medical analysis which is based on genetic algorithm feature for diagnosis of diabetes. Hemant and Pushpavathi [21] involved K-means cluster classification algorithms for predicting diabetes. Santhanam and Padmavathi [22] investigated the dimension reduction of diabetes diagnosis by applying K-means and GA by integrating SVM.

4 Variants of Diabetes Mellitus (DM)

4.1 Type 1 Diabetes (T1DM)

Type 1 diabetes is an insulin-dependent disease, and it is developed when the important core to produce insulin, the beta cells in the pancreas, is damaged, making shortage of insulin which is a vital player in lower the blood sugar. In a lifetime of human being, this may happen at any age but the chances of getting such issues are high in the middle teenage. No remedial measure has so far found to prevent type 1 diabetes. To lead a life, people with type 1 diabetes must have insulin by injection or pump. Intake of Insulin will lead to weight gain which ultimately slows down the diabetes treatment causing some side effects and leads to cardiovascular risk profile and which increases morbidity and mortality.

4.2 Type 2 Diabetes (T2DM)

In early stages of type 2 diabetes, the disease does not show any symptoms, so patients are unaware about diabetes and go undiagnosed for several years [23]. Type 2 diabetes is developed due to improper utilization of insulin given by the Beta cells

and developed as insulin resistance. The risk of developing type 2 diabetes is allied with aging, obesity, family unit history of diabetes gestational diabetes, and lesser physical activity.

4.3 Type 3 Diabetes (T3DM)

Alzheimer's disease (AD), also referred as type 3 diabetes, is a chronic neurodegenerative disease, not directly related to DM, but the recent studies support the fact that DM and AD have a sturdy connecting bond [24, 25]. The study analyzed and revealed the relationship between DM and AD via semantic data mining.

4.4 Gestational Diabetes Mellitus

It refers to glucose tolerance with onset or first recognition during pregnancy.

4.5 Protein-Deficient Diabetes Mellitus (PDDM)

It is clearly the commonest type of diabetes that affects the poorest sections of the population, since it occurs due to protein-deficient pancreatic diabetes (PDPD) in all parts of India and can be detected wherever there is awareness. High incidence of PDDM has been reported from Jabalpur, Madhya Pradesh. There is a striking clinical and epidemiological association of PDDM with protein malnutrition. Associated micronutrient deficiency could cause free radical damage to sensitive B-cells [26].

4.6 Impacts of HbA1C (Long-Term Blood Sugar Level)

Complications of type 2 diabetes are robustly connected with high hemoglobin A1c (HbA1c). Micro-vascular complications include nerve system disorder, problem related to eyes (retinopathy), and neuropathy; while macro-vascular complications are primarily understandable by atherosclerosis and follow-on cardiovascular morbidity and mortality [27, 28]. The relationships between glycated hemoglobin (HbA1c) level in patients with DM and without DM have the future risks of cardiovascular disease and death. The relationship between HbA1c and macro-vascular disease appears to be more complex explored by the clinical study which shows direct connection between them. The clinical analysis of HbA1c test results can reduce the risk of all DM-related disease complications.

5 Clinical Laboratory Test Results (Validation)

5.1 Blood Sugar (Glucose Level)

Maintaining the level of glucose in the blood is a key factor to minimizing the risk of complications from type 1 or type 2 diabetic. Patients' blood sugar can be tested following an overnight fasting in the morning before eating and drinking anything. Physicians may, however, choose to test a Random Blood Sugar (RBS) level that can be drawn anytime. No need of prior fasting.

6 Diabetes-Related Comorbidities

Understanding, analyzing, and evaluating comorbid conditions are a crucial part of health management because comorbidities play a vital role in increasing the risk and diagnostics costs. Comorbidity is the presence of more chronic conditions in patient with prominent disease and occurs frequently among patients with diabetes [29, 30]. Comorbidities of DM- and non-DM-related comorbidities, major and well-known DM comorbidities, and high-frequency comorbid with T1DM are explained in Tables 1 and 2.

6.1 Effects of Comorbidity on HC Utilization

It can be estimated in three ways, which is depicted in Fig. 1.

- (1) The occurrence/incidence of any comorbidity
- (2) The cause of type of comorbidity
- (3) The outcome of specific comorbidity.

The patients with diabetes- and comorbidity-associated issues have significant impact for health care and related costs [30–35]. Multilevel analyses have to be followed to estimate the effects of comorbidity on healthcare exploitation by two levels, namely (1) practice level and (2) patient level.

6.2 Diabetes Retinopathy

People who have diabetes are possible to get the major eye problems like cataracts, glaucoma, and retinopathy. DR is retinopathy, and it only affects people who have had diabetes for a long time period and can result in blindness/loss of vision. Awareness about DR is underprivileged in all sectors of people including patients and care

Table 1 Comorbidities of DM and clinical laboratory results [36]

| | | | |
|--|---|---|---|
| Comorbidities | Cardiovascular diseases (pulmonale, cardiomyopathy, valvular disorder, tachycardia, diseases of arteries and veins) | Hypertension (it is a highly sex-sensitive comorbidity) | Renal failure (nephropathy) |
| | Obesity | Anemia and iron deficiency | Diabetes foot (amputations of leg and foot—foot problems) |
| | Retinopathy (eye disease) | Neuropathy (nerve damages) | Gastropathy |
| | Respiratory (lung) diseases | Parkinson's disease | Mental disorders (controversial CM associations) |
| Clinical laboratory results (abnormal) | HbA1C (hemoglobin) | Glucose serum | Fasting blood sugar (FBS) |
| | Troponin-1 | Creatinine | Albumin |
| | LDL/HDL ratio | Triglycerides | Hematocrit (Hct) |
| | Platelet count (Plt) | Mean platelet volume (MPV) | |
| | Lipase | Ketones | Amylase |
| | Potassium | Clexane | |

Table 2 DM- and non-DM-related comorbidities [37]

| S. No. | DM-related comorbidity | Non-DM-related comorbidity |
|--------|------------------------|----------------------------|
| 1. | Heart diseases | Lung diseases |
| 2. | Stroke | Neurological diseases |
| 3. | Retinopathy | Musculoskeletal diseases |
| 4. | Nephropathy | Depression |
| 5. | Diabetic foot | Cancer |

providers; this is because no study has been done so far in the national level to analyze the menace of diabetes. Only paramedics and some sectors were aware about the risk of diabetes which was a risk factor for retinopathy.

6.3 Diabetic Peripheral Neuropathy

This is common among the diabetes people of type 2 diabetes and have the chances of neuropathy, and therefore, they are prone to have diabetic foot ulcer which is core complication in patients with diabetes and lead to amputation of 15% diabetes patients.

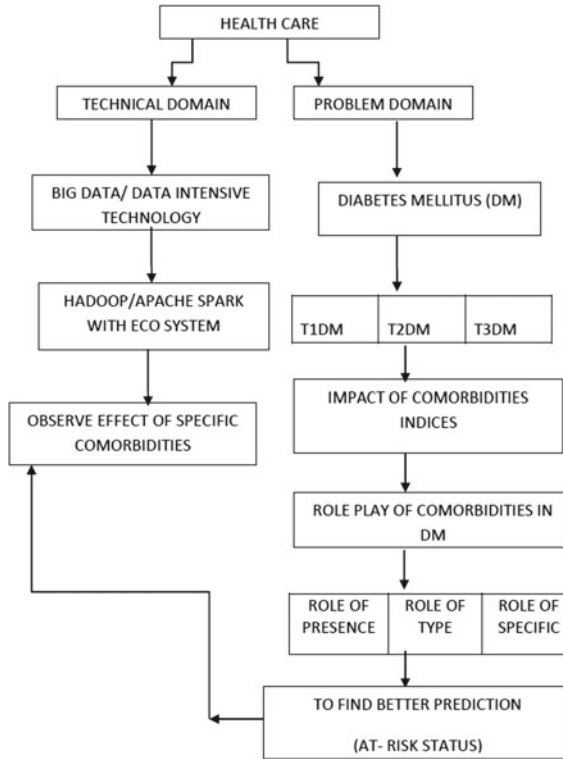


Fig. 1 At-risk prediction flow status in DM comorbidities’ effects through Big Data technology

6.4 Diabetic Nephropathy

It is a leading cause of renal failure resulting in the passage of albumin passes into the urine. This can be avoided by early identification of patients with diabetes and their renal function. This will prevent them to move toward the advanced stage of nephropathy.

6.5 Heart Disease

Diabetes doubles the chances of prone to get a heart disease. Diabetes increases the risk of getting vascular diseases and other cardio-related diseases too.



7 Computational Insights and Interfacing into Diabetes Research with New Parameters (Attributes)

Data mining is a core component in machine learning that arises as a key process providing a very efficient computational insights, where supervised (ANN, SVM, FL, KNN, NB, RF, C5.0) and unsupervised learning algorithms (Clustering) are in existence, which will be more relevant and suitable for medical applications and can be applied for the tasks of classification/clustering/linear regression-enhanced accuracy processes.

By emphasizing a new and evolving risk factors (features) like biomarker selection (HbA1c), metabolic syndrome—metabolism-related pathophysiology, deficiency (imbalance) of some minerals such as magnesium, chromium, and calcium can intensify oxidative stress, intestinal micro-biota (micro-organism—bacteria's role in human body—plays an important role in T2DM risk factor), Hct, Hgb, Plt, and MPV represent risk factors for complications in diabetic patients.

In visualization of prediction, the core is to know about gene, which is studying the human body and its components, population features, risk factors that was already known, therapeutic and medicine history data, and laboratory measurements. Early finding of DM comorbidity by using the influenced (new) interfacing risk factors would definitely open up unique findings into diabetic research (results) when accommodating machine learning library via Big Data computing technology. The abundance of data especially healthcare data in this data centric world forces the technocrats and care providers to move forward toward innovative, economic medical treatment [38].

7.1 Diabetes Medical Datasets

The following are the different categories of diabetes medical datasets, where attributes can easily be accessed;

- DIABCARE—WHO/Europe (quality of health systems)
- PIMA Indian diabetes—UCI repository
- WISCONSIN diabetes registry
- ABBOTT—Diabetes Care Inc.
- DM diagnosis ontology (DDO) and
- Bio-medical bridges diabetes ontology (DIAB).

8 Big Data Technology and Its Prediction Potential—An Unveil

8.1 Analytics

Analytics combine the data by using statistical, mathematical models, programming methodology, problem-solving techniques in resourceful ways. The perceptiveness toward data is to analyze, clean, and identify different patterns.

8.2 Data and Data Science

Data centric world where digital data deposited on digital space is growing in a rapid way ranging from zeta bytes. Data is classified into unordered and ordered data. Data science is a field of study which includes an aspect of data like cleansing, preparation, and analysis.

8.3 Data Analytics and Big Data (BD)

Data analytics is the growing science to explore the raw data and can be done by applying algorithmic techniques for the intention of discovering different formats and thereby forcing to come to a final decision about that data, whereby to derive the insights. (i.e., it means that “Data analysis gives you *an* answer, not *the* answer”).

Big Data, also referred and a part of data-intensive technology, involves data in a high volume and streaming of data with high speed and/or high variety. It is a form of manipulating, modifying, and working on the information which facilitate the improved insights.

8.4 Big Data Analytics Infrastructure (BDAI)

In a general structure, to carry on the BDA and its related task, it needs some specific applications and services. These are collectively called as BDAI. It includes grouping the data with reference to their intrinsic behavior, HDFS, different data analytics tools which may built on Hadoop, Databases/servers SQL, and NOSQL (in-memory data storage).

8.5 BD Analytics (BDA), Types, and BD Technology (Data-Intensive Technology)

Big Data comprising in-memory storage and processing (in-memory execution) elements to gain a “Big Wits” revealed as a tedious task in analytics of BD to fuse data

in a maximum values, called Big Data Analytics. Since data analytics refers to the organized data analysis to transform data into information and knowledge, Big Data technology is a concept of data-intensive computing through machine learning for discovering patterns and exploration of hidden patterns. This technology provides a powerful infrastructure in a distributed environment. There are four types of analytics such as; Predictive, Descriptive, Diagnostic, and Prescriptive.

8.6 Health Big Data

Big Data in health care is a vast platform due to its abundant data, its diversity, nature, arriving speed, etc. Health Big Data includes data in different form to facilitate care providers. The data may be machine generated, data taken by sensor, social media, etc.

8.7 Big Data Properties/Dimensions/Characteristics and Big Data Sources

Recently, it reaches to six levels such as volume, variety, validity, velocity, veracity, and volatility.

Massive data are generated through different sources such as *archives, media, business application, public Web, social media, data storage, and sensor data.*

8.8 Healthcare (HC) Constituents

Recently, all types of people related to research, government sectors, social media, healthcare providers, health analyst, etc., were influenced by Big Data. This can foresee how these groups of people likely to conduct themselves, exhibiting acceptable manners and curtail disagreeable actions. Leveraging Big Data will certainly be a part of the solution to controlling, escalating healthcare costs.

8.9 BDA in HC Methodology: Drafting Pseudo-code (Refer Table 3)

As part of model development in any medical (health care) application, a stage-by-stage execution (pseudo-code) is to be entertained, detailed in an understandable way for the big data handlers.

Table 3 BDA—healthcare—methodology—pseudo-code

```
Begin
{
  Step 1: Defining Model
  Requirements of BDA-HC
  Step 2: Defining Plan
  Drafting the Plan
  Referring the Research gap
  Reason for the problem selection
  BDA-Analytical Approach
  Required Materials
  Step 3: Methodology
  Selection of Attributed
  Dataset collection
  Data conversion
  Selection of BDA Tools/BD platforms
  Defining a workable model
  Algorithms
  Results & Inference
  Step 4: Outcome Analysis
}
End;
```



9 Big Data Platforms/Technologies/Core and Ecosystems (Refer Table 4)

Hadoop architecture (Master/Slave model) consists of Hadoop common and three core components like HDFS, MR, and YARN. Hadoop is concerned with three types, such as Apache Hadoop, RHadoop, and Hortonworks Hadoop. Hadoop MR framework allows massive scalability and processing of extremely large file. MR is the heart of Hadoop where the processing is carried out by assigning the tasks to various clusters. Big Data-driven medicine is an evolving research concept that will enable the discovery of new treatment opinions through the new research thrust on comorbidities of diabetes disorder. All multifaceted attributes related to Big Data are defined in the Big Data ecosystem (BDE) which is capable of managing the developing arena.

Table 4 Big Data ecosystems projects

| Hadoop ecosystem projects | |
|-------------------------------------|--|
| Pig (Pig Latin) | Process text image |
| Impala | SQL query engine designed for BI |
| Sqoop | Data movement to (Hadoop ecosystem)/from RDBMS |
| Hbase (built on Hadoop file system) | Can store diverse data |
| Mahout (ML Library) | Supports various data mining algorithms |
| Oozie | Utilizes the Java runnable platform |
| Hypertable | Used for storing data. This is similar to NOSQL data repository |
| Riak | Smart working ability like ease availability, fault-traces, extendable through distributed NOSQL |
| Apache Hive (To replace MR) | Acts as data warehouse for clusters of Hadoop. Data fetching can be done by HiveQL |
| FlockDB | Database of social media graphs |
| Hibari | Capable of storing high-bandwidth data |
| Apache Flume | Processes large volumes of log files, scalable architecture, and provides reliable solution |
| Kafka | Streaming data ingestion |
| Solr | Enables users to find the data they need |
| Hue | Graphical front end to cluster. Well based user interface for Hadoop |
| Storm | Streaming |
| Avro | Data access |
| Zookeeper | Management (high availability directory) |

9.1 BDA: Platforms

Platforms of BDA are Apache Hadoop, Apache Spark (growing and influencing platform in Big Data), Apache Hive, Apache Flume, Apache Sentry, IBM Infosphere, Apache S4, Twitter's Storm, Dremel (Google's Big Query services), Lambda Architecture, and epic. Hadoop tools and other tools can help and solve real problems;

1. Data storage: Hbase
2. Data integration: Flume, Kafka, and Sqoop
3. Data processing: Spark and MR
4. Data analysis: Hive and Impala
5. Data exploration: Cloudera Search
6. Data security: Sentry (authorization tool providing security for Hadoop).

Note: Pig and Hive are projects built to replace the coding the MapReduce.

9.2 Hive and R

Hive is an technology, extended to Hadoop, namely MR and data storage to end-users. Beginners can write MR programs using this. Hive is on crest of Hadoop and provides HiveQL. HiveQL is normal SQL model. All are stored as Hadoop Distributed File system repository. The HDF core processes the HiveQL statements. But before processing, the HiveQL statements must be placed as map reduce jobs on a Hadoop Cluster.

9.3 NOSQL (In-Memory Data Base Storage)

There are four NOSQL types, such as key value pair DB, document store DB, column store DB, and graph-based DB. Here, column store and document store DB are more suitable for medical applications.

10 Conclusion

While visualizing the new and evolving parameters like HbA1C, Hct, Hgb, Plt, Imbalance (deficiency) minerals, and micro-biota, it was suggested that these parameters be considered as the potential risk factors. These parameters plays a core role in human body in getting T2DM of diabetes mellitus risk factors in the line of data-driven comorbidities indices as it is known to increase the risk factors. These parameters are to be considered along with Data-Intensive Technology's strengths' so as to disclose or find or extract hidden patterns delivery on DM patients' care prediction

research. This analysis part of our research study identified (through literatures) the most importance in the inclusion part of aforesaid risk factors in the future investigation, would cover the at-risk prediction status, which was not yet touched as a problem domain in computer application. Hence, this analysis (unique findings) delivery will definitely impact the researchers and also an eye-opener to enhance the research potentials on the prominent cause of disorder, called diabetes mellitus. A detailed outline and role behind Big Data potentials especially in DM would deliver the more accuracy results (to target at-high-risks) to reduce mortality rate in the developing country like India.

References

1. The Lancet. (2011). The diabetes pandemic. *The Lancet*, 378(9786), 99. [https://doi.org/10.1016/s0140-6736\(11\)61068-4](https://doi.org/10.1016/s0140-6736(11)61068-4). PMID: 21742159.
2. Ekoé, J. M., Rewers, M., Williams, R., & Zimmet, P. (2008). *The epidemiology of diabetes mellitus*. New Jersey, USA: Wiley.
3. Haffner, S. M., Lehto, S., Rönnekaa, T., Pyörälä, K., & Laakso, M. (1998). Mortality from coronary heart disease in subjects with type 2 diabetes and in non-diabetes subjects with and without prior myocardial infarction. *New England Journal of Medicine*, 339, 229–234. PMID: 9673301.
4. Almdal, T., Scharling, H., Jensen, J. S., & Vestergaard, H. (2004). The independent effect of type 2 diabetes mellitus on ischemic heart disease, stroke and death. *Archives of Internal Medicine*, 164(13), 1422–1426. PMID: 15249351.
5. Anderson, R. J., Freedland, K. E., Clouse, R. E., & Lustman, P. J. (2001). The prevalence of comorbid depression in adults with diabetes. *Diabetes Care*, 24(6), 1069–1078. PMID: 11375373.
6. Engum, A. (2007). The role of depression and anxiety in onset of diabetes in a large population-based study. *Journal of Psychosomatic Research*, 62(1), 31–38. PMID: 17188118.
7. Fong, D. S., Aiello, L., Gardner, T. W., King, G. L., Blankenship, G., Cavallerano, J. D., et al. (2004). Retinopathy in diabetes. *Diabetes Care*, 27, 584–587.
8. Lago, R. M., Singh, P. P., & Nesto, R. W. (2007). Diabetes and hypertension. *Nature Clinical Practice Endocrinology & Metabolism*, 3, 667.
9. Sharmila, K., & Vethamanickam, D. S. (2015). Survey on data mining algorithm and its application in healthcare sector using Hadoop platform. *IJETAE*, 5(1).
10. Saravana kumar, N. M., Eswari, T., Sampath, P., & Lavanya, S. (2015). Predictive methodology for diabetic data analysis in Big Data. *Procedia Computer Science*, 50, 203–208.
11. Srivastava, A. K., Kumar, C., & Mangla, N. (2016). Analysis of diabetic dataset and developing prediction model by using Hive and R. *IJST*, 9(47).
12. Sadhana, S. S., & Shetty, S. (2014). Analysis of diabetic data set using Hive and R. *IJETAE*, 4(7).
13. Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. *Computational and Structural Biotechnology Journal*. <https://doi.org/10.1016/j.csbj.2016.12.005>.
14. Alic, B., Gubeta, L., & Badnjevic, A. (2017). Machine learning techniques for classification of diabetes and cardiovascular disease. In *IEEE, MECO*, June 2017.
15. Sabibullah, M. (2012). Prognostic neural network model for diabetic risks prediction. In *Proceedings of IEEE International Conference on Emerging Trends in Science, Engineering and Technology* (pp. 392–395).

16. Farran, B., Channanath, A. M., Behbehani, K., & Thanaraj, T. A. (2013). Predictive models to assess risk of type 2 diabetes, hypertension and comorbidity: Machine-learning algorithms and validation using national health data from Kuwait—A cohort study. *BMJ Open Journal*. <https://doi.org/10.1136/bmjopen-2012-002457>.
17. Vijayan, V. V., & Anjali, C. (2015). Decision support systems for predicting diabetes mellitus—A review. In *IEEE, Proceedings of 2015 Global Conference for Communication Technologies (GCCT 2015)*.
18. Srikanth, P., & Deverapalli, D. (2016). A critical study of classification algorithms using diabetes diagnosis. In *IEEE 6th International Conference on Advanced Computing*. <https://doi.org/10.1109/iacc.2016.54>.
19. Patil, B. M., Joshi, R. C., & Toshniwal, D. (2010). Association rule for classification of type-2 diabetes patients. In *IEEE, 2010 Second International Conference on Machine Learning and Computing*. <https://doi.org/10.1109/icml.c.2010.67>.
20. Aishwarya, S., & Anto, S. (2014). A medical decision support system based on genetic algorithm and least square support vector machine for diabetes disease diagnosis. *International Journal of Engineering Sciences & Research Technology*, 3(4), 4042–4046. ISSN 2277-9655.
21. Hemant, P., & Pushpavathi, T. (2012). A novel approach to predict diabetes by cascading clustering and classification. In *Computing Communication & Networking Technologies (ICCCNT): Third International Conference*, July 2012.
22. Santhanam, T., & Padmavathi, M. S. (2015). Application of K-means and genetic algorithms for dimension reduction by integrating SVM for diabetes diagnosis. *Elsevier ScienceDirect*. <https://doi.org/10.1016/j.procs.2015.03.185>.
23. Harris, M. I., Klein, R., Welborn, T. A., & Knudman, M. W. (1992). Onset of NIDDM occurs at least 4-7 yr before clinical diagnosis. *Diabetes Care*, 15, 815–819.
24. De la Monte, S. M., & Wands, J. R. (2018). Alzheimer's disease is type 3 diabetes evidence reviewed. *Journal of Diabetes Science and Technology*, 2(6), 1101–1113.
25. Narasimhan, K., Govindasamy, M., Gauthaman, K., Kamal, M. A., Abuzenadeh, A. M., Al-Qahtani, M., et al. (2014). Diabetes of the brain: Computational approaches and interventional strategies. *CNS & Neurological Disorders-Drug Targets (Formerly Current Drug Targets-CNS & Neurological Disorders)*, 13(3), 408–417, 2014.
26. Tripathy, B. B., & Samal, K. C. (1993). Protein deficient diabetes mellitus (PDDM) in India. *IJDDC*, 13, 3–13.
27. Fowler, M. J. (2008). Micro vascular and macro vascular complications of diabetes. *Clinical Diabetes*, 26(2), 77–82.
28. Stolar, M. (2010). Glycemic control and complications in type 2 diabetes mellitus. *The American Journal of Medicine*, 123(suppl 3), S3–S11.
29. Feinstein, A. (1967). *Clinical judgement*. New York: The Williams & Wilkins Company.
30. Beckman, J. A., Creager, M. A., & Libby, P. (2002). Diabetes and atherosclerosis: Epidemiology, pathophysiology, and management. *JAMA*, 287, 2570–2581. <https://doi.org/10.1001/jama.287.19.2570>.
31. Egede, L. E., Zheng, D., & Simpson, K. (2002). Comorbid depression is associated with increased health care use and expenditures in individuals with diabetes. *Diabetes Care*, 25, 464–470.
32. Van den Akker, M., Buntinx, F., Metsemakers, J. F., Roos, S., & Knottnerus, J. A. (1998). Multimorbidity in general practice: Prevalence, incidence, and determinants of co-occurring chronic and recurrent diseases. *Journal of Clinical Epidemiology*, 51, 367–375. [https://doi.org/10.1016/s0895-4356\(97\)00306-5](https://doi.org/10.1016/s0895-4356(97)00306-5).
33. Black, S. A. (1999). Increased health burden associated with comorbid depression in older diabetic Mexican Americans. Results from the hispanic established population for the epidemiologic study of the elderly survey. *Diabetes Care*, 22, 56–64.
34. Gijzen, R., Hoeymans, N., Schellevis, F. G., Ruwaard, D., Satariano, W. A., & van den Bos, G. A. (2001). Causes and consequences of comorbidity: A review. *Journal of Clinical Epidemiology*, 54, 661–674. [https://doi.org/10.1016/S0895-4356\(00\)00363-2](https://doi.org/10.1016/S0895-4356(00)00363-2).

35. Westert, G. P., Satariano, W. A., Schellevis, F. G., & van den Bos, G. A. (2001). Patterns of comorbidity and the use of health services in the Dutch population. *European Journal of Public Health, 11*, 365–372. <https://doi.org/10.1093/eurpub/11.4.365>.
36. Klimek, P., Kautzky-Willer, A., Chmiel, A., Schiller-Frühwirth, I., & Thurner, S. (2015). Quantification of diabetes comorbidity risks across life using nation—Wide claims data. *PLOS Computational Biology*. <https://doi.org/10.1371/journal.pcbi.1004125>.
37. Struijs, J. N., Baan, C. A., Schellevis, F. G., Westert, G. P., & Van Den Bos, G. A. (2006). Comorbidity in patients with diabetes mellitus: Impact on medical health care utilization. *BMC Health Services Research*. <https://doi.org/10.1186/1472-6963-6-84>.
38. Yensen, J., & Naylor, S. (2016). The complementary iceberg tips of diabetes and precision medicine. *Journal of Precision Medicine, 3*, 21–39.

Interpreting the Objective Outcome of the Proposed Misuse Case Oriented Quality Requirements (MCOQR) Framework Metrics for Security Quantification



Ajeet Singh Poonia, C. Banerjee, Arpita Banerjee and S. K. Sharma

Abstract A number of tools, techniques, methods, methodology, and standards are available to quantify the security aspect of software during its development and after it has been implemented. But the interpretation and analysis of the quantified security metrics thus obtained may be difficult for the software development team. Proper and comprehensive interpretation and analysis of quantified security metrics are essential to specify correct security requirements during the requirements engineering phase of SDLC which may result in more secured software. This research work shows how the proposed Misuse Case Oriented Quality Requirements (MCOQR) framework metrics may be used to provide identification, definition, interpretation, and analysis of security metrics during the requirements engineering phase of software development process. The authors also discuss the various primary outcomes that may be obtained using the proposed MCOQR framework metrics using the industry accepted standards like Common Vulnerability Scoring System (CVSS), Common Vulnerability Enumeration (CVE), and Common Weakness Enumeration (CWE). The work proposed is an extension of Misuse Case Oriented Quality Requirements (MCOQR) framework metrics and includes software application-specific database. The study also highlights the areas where future research work can be carried out to further strengthen the entire software system during the software development process.

A. S. Poonia

Government College of Engineering and Technology, Bikaner, India

e-mail: pooniaji@gmail.com

C. Banerjee (✉)

Amity University Rajasthan, Jaipur, India

e-mail: chitreshh@yahoo.com

A. Banerjee

St. Xavier's College, Jaipur, India

e-mail: arpitaa.banerji@gmail.com

S. K. Sharma

Modern Institute of Technology and Research Center, Alwar, India

e-mail: sharmasatyendra_03@rediffmail.com

Keywords Security metrics · Misuse cases · CVSS · CVE · MCOQR metrics

1 Introduction

Due to the emergent information communication technology (ICT)-centric economy, there seems to be a requirement for the development of a security metrics which follows the SMART concept; i.e., a security metrics which is specific in nature can be efficiently measured, is easily attainable, can be made repeatable, and is time dependent. Moreover, the security metrics should be totally integrated into the software development process right from the beginning for delivery of secured software [1].

The basic objective of security which enforces that the secured software thus development cannot be intentionally undermined or force to fail, remains correct and predictable despite of its dependability compromise, continues to operate correctly regardless of various attacks, either prevent or limits the damage caused due to the attacks. In short, the secured software should be attack-resistant, attack-tolerant, and attack-resilient [2, 3].

Although a good number of security requirements engineering techniques and approaches are proposed by the researchers and adopted by the industry, still the issue is in its infancy stage and requires refinement and realignment from the stakeholder point of view. The proposed techniques and approaches are too broad in nature to be made practically implementable by the small-scale and medium-scale software development industry. Further, they also need high level of expertise and technicalities in terms of skilled manpower. Also, they advocate more about the standards and protocols to be adopted for the implementation of security during the SDLC [4, 5].

According to a recent study, it was concluded that when the application is built and practically used in the market and reports heavy security breach on daily basis, to improve upon the security aspect becomes a difficult and costly affair. Moreover, it was also found that the return on investment when the security aspect was considerably addressed during the early stages of software development process was at its highest as compared to an application development which neglected and ignored the implementation of security during the development process [6].

The software industry, specially the small and medium scale, is now motivated toward the implementation of security during the early stages of software development process. But they need a comprehensive yet easy to implement a security approach which is inclined more toward the learning and analysis aspect of security implementation during the SDP while creating awareness among its stakeholders. Further, since the security is an intangible and non-functional property of software, hence its quantification is also necessary to understand and further improve the security process [7–9].

In this research work, a comprehensive workflow is shown which helps in interpreting the objective outcome of the proposed Misuse Case Oriented Quality Requirements (MCOQR) framework metrics for security quantification. Apart from Sect. 1

which mainly focuses on introduction, the rest of the research work contains the following: Sect. 2 presents the proposed algorithm for creation of misuse case modeling tree with count, scoring, and ranking, Sect. 3 discusses the implementation mechanism of the proposed algorithm, Sect. 4 presents the results and discussion, whereas in Sect. 5 the conclusion and future research work is given.

2 Proposed Work

A good number of guidelines, methods, and processes are available in the market for the security implementation during the software development life cycle, but they are too broad in nature and their practical implementation by the software firms is very difficult. Even if the software firms adopt any of these available security guidelines, methods, or processes, it is impossible to produce 100% secured software and there is a reasonable probability that the software developed will be introduced in the market with both known and unknown vulnerabilities which may be exploited by the attacker to harm the system.

However, some comprehensive and improvised method or process may be developed which if adopted may try to minimize the risk associated with the exploitation of these vulnerabilities by the attacker and the software is able to operate while maintaining confidentiality, integrity, and availability aspect. There are a number of ways to minimize the security risk during the development of the software application, but if the proposed method is easy and flexible to adopt, having involvement of almost all the stakeholder, well synchronized with some available security standards validated by the industry and provides some tangible figures as indicators and estimators for

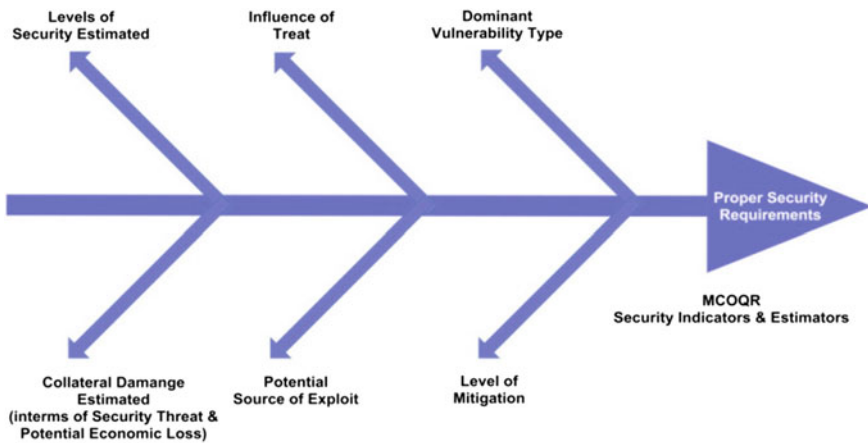


Fig. 1 MCOQR security indicators and estimators

further analysis then the software firm may be readily available to implement it to build secured software (Fig. 1).

One such framework metrics proposed is Misuse Case Oriented Quality Requirements (MCOQR) framework metrics which is based on the identification of vulnerabilities using industry standards like CVSS, CVE, and CWE, correlating these vulnerabilities with associated misuse cases and providing a classification of these identified vulnerabilities which are presented using the proposed metrics and provides five indicators and estimators to judge the level of implementation of security during the requirements engineering phase of software development process. These five indicators and estimators may aid the security requirements team to specify a comprehensive set of security requirements for the development of secured software.

3 Implementation Mechanism

The misuse case modeling technique could prove to be a potential contender for identifying, defining, quantifying, and specifying security requirements during the early phases of software development process, i.e., from the requirements elicitation process. Keeping the above explanation and conclusion in mind, we propose Misuse Case Oriented Quality Requirements (MCOQR) framework which is a risk-based security requirements engineering framework and which may help the security requirements engineering team for quantification of software security (Fig. 2).

The results thus obtained by the implementation of the proposed framework metrics may act as an indicator and estimator for the security requirements engineering team to quantify security of software being in the development process. The proposed work may enable the security requirements engineering team to access the security risk easily and efficiently well in advanced, and some proactive approach may be designed and implemented in the form of mitigation mechanism for securing the software application before it is developed.

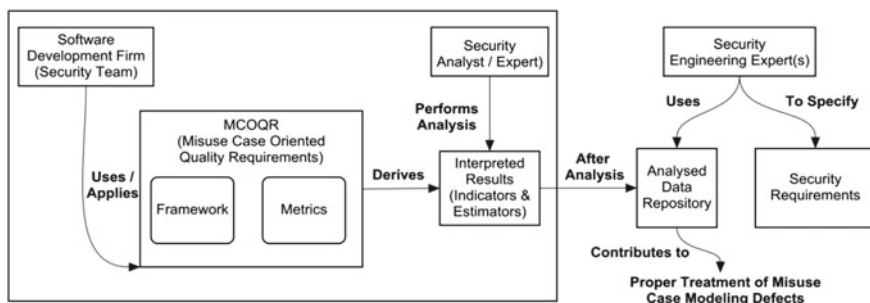


Fig. 2 Implementation mechanism



4 Results and Validation

The proposed algorithm was applied to a real-life project from industry (on the request of the company, identity is concealed), and the final result of security assessment is calculated as per prescribed implementation mechanism.

Then, the level of security assurance is compared with the other project's security assurance in which the proposed algorithm was not applied. The study shows that the security which is an intangible entity can be transformed into a tangible entity which may help the software development team to quantify the security aspect of software during the security requirements engineering phase of software development process, thereby minimizing the risk caused during the implementation of software in the market. Due to the page limit constraint, we are not providing the details of validation results in this paper; we will discuss in our next paper.

5 Conclusion and Future Work

We propose MCOQR-based metrics and statistical metrics derived to predict five indicators and estimators which may be used by the security team and requirements engineering team to specify security requirements during the requirements engineering phase of software development process. The proposed framework process and five parameters as predicted indicators and estimators have been verified and validated by the security experts from various government and non-government organizations.

Distinction between the functional and the non-functional security requirements of a system is subtle. One security feature may be adopted as function requirement for the programmer, and the same security feature may be categorized as non-functional requirement for the user of the system. The study is to be carried out to find this fine distinct line and gap which is a good contender for future research work.

References

1. Morrison, P., Moye, D., & Williams, L. A. (2014). *Mapping the field of software security metrics*. Department of Computer Science: North Carolina State University.
2. Banerjee, C., & Pandey, S. K. (2009). Software security rules. *SDLC Perspective*. *arXiv preprint*.
3. Banerjee, C., Banerjee, A., & Pandey, S. K. (2016). MCOQR (Misuse case-oriented quality requirements) metrics framework. In *Problem solving and uncertainty modeling through optimization and soft computing applications* (pp. 184–209). IGI Global.
4. Rehman, S., & Gruhn, V. (2017, September). Security requirements engineering (SRE) framework for cyber-physical systems (CPS): SRE for CPS. In *New Trends in Intelligent Software Methodologies, Tools and Techniques: Proceedings of the 16th International Conference SoMeT_17* (Vol. 297, p. 153). IOS Press.
5. Braude, E. J., & Bernstein, M. E. (2016). *Software engineering: Modern approaches*. Waveland Press.

6. Mead, N. R. (2006). Identifying security requirements using the security quality requirements engineering (SQUARE) method. *Integrating Security and Software Engineering*, 44–69.
7. Banerjee, C., Banerjee, A., & Murarka, P. D. (2014). Evaluating the relevance of prevailing software metrics to address issue of security implementation in SDLC. *International Journal of Advanced Studies in Computers, Science and Engineering*, 3(3), 18.
8. Karim, N. S. A., Albuolayan, A., Saba, T., & Rehman, A. (2016). The practice of secure software development in SDLC: an investigation through existing model and a case study. *Security and Communication Networks*, 9(18), 5333–5345.
9. Raj, G., Singh, D., & Bansal, A. (2014, September). Analysis for security implementation in SDLC. In *2014 5th International Conference on Confluence the Next Generation Information Technology Summit (Confluence)* (pp. 221–226). IEEE.

A Comparative Performance Study of Machine Learning Algorithms for Sentiment Analysis of Movie Viewers Using Open Reviews



Dilip Singh Sisodia, Shivangi Bhandari, Nerella Keerthana Reddy and Abinash Pujahari

Abstract The Internet facilitated the easy access to public opinions and reviews for any product or services. The collective opinions of people are significantly helpful for making decisions about any product or services. Movies are one of the most captivating pass times of the modern world on which people like to give their opinion/review. Movie reviews are personal opinions or comments shared via social media tool by a common viewer who has watched the movie. It provides an opportunity to know the outreach and response of the viewers to any film. Movie reviews influence the decision of prospective viewers as well as help producers, directors, and other stakeholders for improving the quality aspect of the movie. Therefore, movie reviews play an important role in sentiment analysis of target viewers. In this research work, the sentiment analyses of viewers based on movie reviews using machine learning methods are discussed. The raw movie reviews are collected, and after performing preprocessing, features are extracted using bag of words, TF-IDF, bigram methods from text reviews. Various machine learning techniques including Naive Bayes classifier, Support Vector Machine, Decision trees, and ensemble learners are used with different feature extraction schemes to obtain a sentiment analysis model for positive or negative polarity in the movie review data sets. The performance of learners based sentiment analysis model is evaluated using accuracy, precision, recall, and f-measures. The objective of this research is to find the best classifier to test the reviews of movies given out by people so that we would know the overall general opinion of the audience. It is concluded that the set of classifiers can be used collaboratively to get effective results. Changes can be made from the very algorithmic level of the classifiers to gain better performance in the domain of study.

D. S. Sisodia (✉) · S. Bhandari · N. K. Reddy · A. Pujahari
National Institute of Technology Raipur, Raipur, India
e-mail: dssisodia.cs@nitrr.ac.in

S. Bhandari
e-mail: shivangibhandari1@gmail.com

N. K. Reddy
e-mail: nkeerthanareddy1@gmail.com

A. Pujahari
e-mail: abinash.pujahari@gmail.com

© Springer Nature Singapore Pte Ltd. 2020
M. Pant et al. (eds.), *Performance Management of Integrated Systems and its Applications in Software Engineering*, Asset Analytics,
https://doi.org/10.1007/978-981-13-8253-6_10

Keywords Sentiment analysis system · Movie reviews · Movie opinion · Sentiment analysis model · Machine learning algorithm

1 Introduction

We all tend to take others' opinions before doing even trivial matters like shopping or important issues like voting for an election. In older days, neighbors, friends, and families used to give insights and opinions to us. But these days, the Internet is the biggest source of information for all our needs [1].

Movies are one of the most captivating pass times of the modern world on which people like to give their opinion/review. These reviews mostly appearing on social media like Facebook, Twitter [2], reviewing blogs, provide an opportunity to know the outreach and response of the viewers to any film. These reviews are not done by the film critics or film gurus but by the common people. They contain completely raw and unbiased (most of the time) opinion on the said film [3]. For the movie makers or the viewers who have not watched the film yet, the summary of all reviews, i.e., whether they are positive or negative [4], helps them decide the outcome. But, trying to get to a conclusion by manually reading each review would prove to be a very tedious task for a single person or even for an organization. Hence, we need a cumulative summary of all the reviews.

Hence, the job of any sentiment analysis system is to analyze, summarize, and classify the data automatically. The job of sentiment analysis may be done at different levels like at the document level, sentence level, word level, or aspect level. While considering the sentiment analysis at document level, it assumes that reviews have an opinion about single entity [5]. Sentence-level analysis determines opinion for independent sentence or review (subjectivity analysis). However, both of these analyses fail to find the exact like or dislike of peoples. Thus, aspect level is one area of sentiment analysis which can provide accurate results for sentiment analysis system [6].

This paper makes use of two data sets which are of around 25,000 and 1000 reviews, respectively, consisting of different movies given out on the Internet and then randomly dividing the entire data into different sets which are used as inputs to various machine learning algorithms. These machine learning algorithms work on some selected features of the data which are extracted using natural language processing techniques. We compared the performance of different machine learning algorithms for the given selected features.

2 Related Works

Sentiment analysis is the process of analyzing the views of persons toward any items or service. A lot of research has been done in this domain [7]. Identifying

subjective sentence of any person is a difficult task with the perspective of sentiment analysis [3], although subjective analysis system is responsible for the sentiment analysis performance increase. Riloff et al. proposed bootstrapping process [8] for subjectivity classification. They used two classifiers where one classifier searches for the subjective sentences in the data set and the other searches for the objective sentences.

H. Yu et al. proposed subjectivity classification [9] in terms of sentence similarity and the Naive Bayes classifier. In this paper, they have used SIMFINDER system to measure the similarity in sentences in terms of words, phrases, and WordNet Synsets. Pang et al. [10] made use of learning techniques like Bayesian and SVM algorithm to classify the movie reviews.

P. Bhoir et al. proposed two different methods [4] like obtaining the subjectivity of sentences and then used a rule-based system which is used to find feature–opinion pair. Then, the representation of the opinion is done using the two methods. First, the proposed system used SentiWordNet approach to get the orientation of extracted opinion and then it uses a method which is based on lexicon consisting list of positive and negative words.

Similar kind of work has been done in other languages too. For example, sentiment analysis of French movie reviews has been done in [11] where experiments are conducted by converting French text reviews into English text.

3 Methodology

To evaluate various machine learning techniques for the sentiment analysis, the data set is obtained online, which are described in detail in the experiment section. We then use the following methods to find the accuracy of the built model. The summary of the adopted methodology is shown in Fig. 1.

3.1 Prerequisite

The first and foremost step is collecting the data set. On this data set, we perform cleaning, where HTML tags and stop words are removed, data is split into lower cases only, and a list of required words is created among other functions. Then, we decide the various features based on which the classification is performed. These features are then extracted based on the code. A feature vector represents each review.

All the feature vectors of the training samples are passed as input to the classifier along with the class label they belong to, and thus, the classifiers undergo learning. After the learning of the classifiers is completed, the feature vectors of the testing samples are sent as input to the models and are labeled by the classifiers.

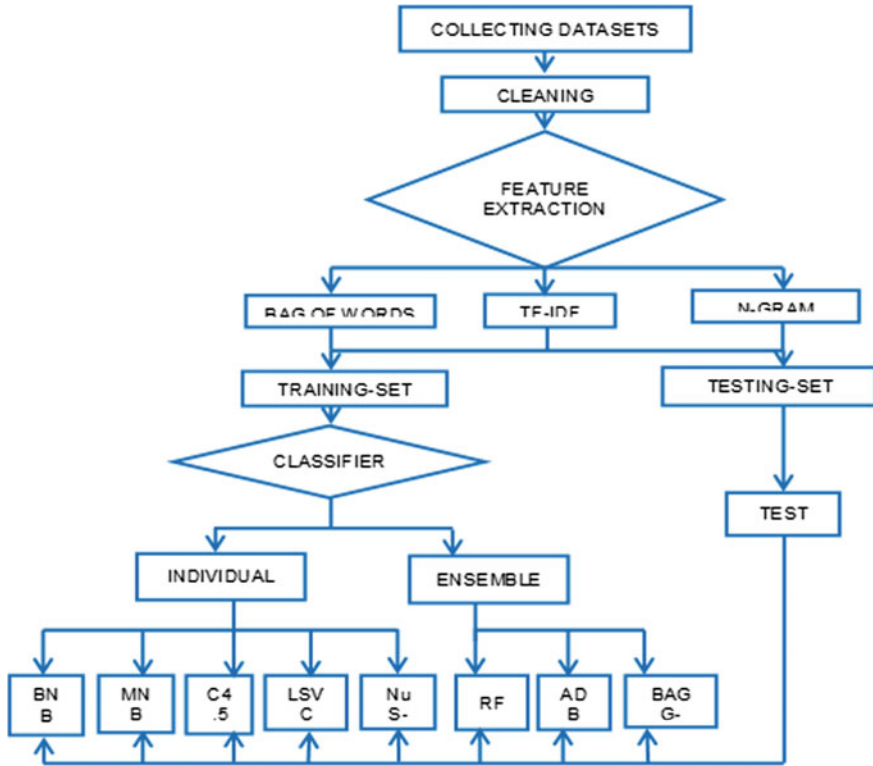


Fig. 1 Block diagram of methodology

3.2 Feature Extraction

Feature extraction is a process where features in the data that contribute most to the prediction variable or output are automatically selected. Three types of features extraction are used here: bag of words, TF-IDF, and bigrams.

Bag of words [12] is a sparse vector of the occurrence of the count of words in the document. It treats the words as images, and the images are treated as documents. This model is good for assessing small data sets.

The term frequency–inverse document frequency [13] is a statistic that shows how important a word is to a document to collect its corpus. It increases the proportionality of a word that occurs frequently but is offset when it occurs more number of times in the corpus. According to IDF, the relevance of a word could increase if it is mentioned less frequently.

An n-gram is a contiguous sequence of n items from a given sequence of speech or text. In our study, we used bigram [9] which is defined as the sequence of two elements from a string of tokens. It is an n-gram where $n = 2$. The bigrams help



to provide the probability of a token when the probability of the previous token is already known to us.

4 Learning Algorithms and Performance Measures

Text classifiers which are based on machine learning algorithms are a kind of supervised machine learning. Here, the classifiers are needed to be trained on some training data before they are applied to the actual set (classification task). The training data is an extracted part of the data which are hand labeled manually. Classifiers are first trained so that they can be used to test actual data suitably. Three individual learners and three ensemble learners were used for carrying out our study which is available in the scikit-Learn package [14] for Python. Bayes' Theorem can be written as:

$$P(C|t_i) = \frac{P(C) \times P(t_i|C)}{P(t_i)} \quad (1)$$

Equation 1 provides a way to calculate the posterior probability, i.e., $P(C|t_i)$, from prior probability $P(C)$, with likelihood $P(t_i|C)$ and $P(t_i)$. Multinomial Naive Bayes is a variant, which is composed more of text documents and explicitly models the word counts and adjusts the underlying calculations to deal with it. The multinomial model [15] caught word frequency data in records and computed class probabilities.

In the Bernoulli document model, every review is represented by a feature vector which will be getting the value one if a specific feature is there in the review and 0 if the feature is not present in the review [16]. Decision tree learning uses a decision tree as a predictive model which maps observations about an item to conclusions about the item's target value [16]. C4.5 is an algorithm used to generate a decision tree developed in [17].

Support vector machines (SVMs) are that examine information and perceive designs and classify data by transforming it into higher dimension using a kernel function and then finding the best hyperplane that separates the patterns of one class from those of the other class [18]. They have been utilized as a part of a wide assortment of uses such as text classification [19], facial expression recognition [18], gene analysis, and much more.

Among the ensemble classifiers, in random forest (also known as random subspace) many individuals, unpruned decision trees [20] are used. Bagging is a bootstrap ensemble method that creates individuals for its ensemble by training each classifier on a random redistribution of the training set. In AdaBoost, iterative process to improve simple boosting process is used. It focuses on the patterns that are hard to classify.

4.1 Performance Evaluation Measures

The evaluation metrics provide greater insight into the performance characteristics of the classifier. Table 1 represents the confusion matrix, which is one of the performance measures used to measure the precision and accuracy [21].

Accuracy (A): It is commonly used as a measure for categorization techniques. Accuracy values are however much less reluctant to variations in the number of correct decisions than precision and recall. Equation 2 will find out the accuracy of the prediction.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{Total No. of Samples}} \quad (2)$$

Precision (P): It is the conditional probability that document 'd' is classified under a class (c_i). It measures the ability of classifiers to place a sample document under the correct class as opposed to all documents which are placed in that class, both correct and incorrect. Equation 3 is used for calculating precision.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (3)$$

Recall (R): A decision is taken to determine the probability whether document d should be classified under class (c_i). Equation 4 is used for calculating recall.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (4)$$

F-Measure (F): Precision and recall are combined to produce a single metric known as F-measure, which is the weighted harmonic mean of precision and recall. Equation 5 below is used to find out F-measure.

$$F\text{-Measure} = 2 \times \frac{P \times R}{P + R} \quad (5)$$

Table 1 Confusion matrix

| | | Predicted class | |
|--------------|--|---------------------|---------------------|
| Actual class | | Yes | No |
| Yes | | True positive (TP) | True negative (TN) |
| No | | False positive (FP) | False negative (FN) |

5 Experimental Results and Discussion

The experiments are performed using two data sets: One is from Kaggle platform which consists of 25,000 reviews [22]. The other data set prepared by collecting around 1000 Bollywood movies reviews from the year 2013–2016 belonging to different genres [23]. Both data sets are divided into training and testing sets. After performing preprocessing, first data set (DS-1) consists of 20,000 reviews while second data set (DS-2) consists of 900 reviews. Table 2 describes the summary of training and testing data sets used for the experimental verification.

By using the data sets mentioned above, we have validated both the data set using different features and classifiers using the Weka tool. We have done separate experiments for the two data sets. The following tables (Tables 3, 4, 5, and 6) represents the different performance values obtained by using different classifiers as well as different feature extraction techniques.

We started with Naive Bayes classifier. From Table 3, BNB classifier provides better performance as compared to MNB classifier for all the three feature extraction techniques for DS1. In feature point of view, bigram is performing better for DS1. For DS2, BNB is performing better for the entire three feature extraction techniques, whereas bag-of-word feature extraction technique is performing better than others for DS2.

Table 2 Data set used for experiment

| Name | Train data positive negative | | Total | Test data positive negative | | Total |
|------------------|------------------------------|--------|--------|-----------------------------|------|-------|
| | | | | | | |
| Data set 1 (DS1) | 9972 | 10,028 | 20,000 | 2528 | 2472 | 5000 |
| Data set 2 (DS2) | 578 | 322 | 900 | 62 | 38 | 100 |

Table 3 Performance metrics: Naive Bayes classifiers

| Classifiers | | Feature | A | P | R | F |
|-------------|-----|--------------|------|------|------|------|
| DS1 | MNB | Bag of words | 0.84 | 0.84 | 0.84 | 0.84 |
| | | TF-IDF | 0.84 | 0.84 | 0.82 | 0.85 |
| | | Bigrams | 0.85 | 0.85 | 0.86 | 0.86 |
| | BNB | Bag of words | 0.84 | 0.85 | 0.84 | 0.85 |
| | | TF-IDF | 0.84 | 0.85 | 0.83 | 0.85 |
| | | Bigrams | 0.86 | 0.85 | 0.87 | 0.86 |
| DS2 | MNB | Bag of words | 0.80 | 0.94 | 0.84 | 0.89 |
| | | TF-IDF | 0.81 | 0.94 | 0.84 | 0.88 |
| | | Bigrams | 0.76 | 0.94 | 0.80 | 0.86 |
| | BNB | Bag of words | 0.90 | 0.97 | 0.92 | 0.95 |
| | | TF-IDF | 0.90 | 0.97 | 0.93 | 0.95 |
| | | Bigrams | 0.86 | 0.98 | 0.87 | 0.92 |

Table 4 Performance metrics: decision tree C4.5 classifier feature

| Feature | DS1 | | | | DS2 | | | |
|--------------|------|------|------|------|------|------|------|------|
| | A | P | R | F | A | P | R | F |
| Bag of words | 0.72 | 0.73 | 0.70 | 0.71 | 0.75 | 0.96 | 0.76 | 0.85 |
| TF-IDF | 0.71 | 0.73 | 0.70 | 0.71 | 0.72 | 0.96 | 0.73 | 0.83 |
| Bigrams | 0.72 | 0.74 | 0.71 | 0.72 | 0.73 | 0.97 | 0.73 | 0.83 |

Table 5 Performance metrics: support vector classifiers

| Classifier | | Feature | A | P | R | F |
|------------|-------|--------------|-------|-------|-------|-------|
| DS1 | SVC | Bag of words | 0.839 | 0.819 | 0.875 | 0.846 |
| | | TF-IDF | 0.839 | 0.820 | 0.871 | 0.846 |
| | | Bigrams | 0.845 | 0.824 | 0.882 | 0.852 |
| | LSVC | Bag of words | 0.829 | 0.833 | 0.827 | 0.830 |
| | | TF-IDF | 0.828 | 0.833 | 0.826 | 0.829 |
| | | Bigrams | 0.837 | 0.846 | 0.828 | 0.837 |
| | NuSVC | Bag of words | 0.871 | 0.865 | 0.882 | 0.873 |
| | | TF-IDF | 0.870 | 0.865 | 0.880 | 0.869 |
| | | Bigrams | 0.874 | 0.867 | 0.886 | 0.876 |
| DS2 | SVC | Bag of words | 0.910 | 0.92 | 0.989 | 0.953 |
| | | TF-IDF | 0.910 | 0.920 | 0.989 | 0.948 |
| | | Bigrams | 0.890 | 0.911 | 0.976 | 0.947 |
| | LSVC | Bag of words | 0.782 | 0.938 | 0.817 | 0.873 |
| | | TF-IDF | 0.773 | 0.935 | 0.820 | 0.876 |
| | | Bigrams | 0.821 | 0.941 | 0.860 | 0.898 |
| | NuSVC | Bag of words | 0.831 | 0.941 | 0.870 | 0.905 |
| | | TF-IDF | 0.829 | 0.943 | 0.876 | 0.910 |
| | | Bigrams | 0.851 | 0.933 | 0.903 | 0.918 |

Next classifier we evaluated is the decision tree (C 4.5) classifier. Table 4 describes the results obtained using both data sets. For DS1, bigram feature selection is performing better as compared to the other two techniques, whereas for DS2 bag of words is performing better as compared to others.

Next, we evaluated different support vector classifiers. Table 5 represents the different performance metrics obtained by applying different support vector classifiers as well as different feature extraction techniques. It can be seen from the table that NuSVC classifier is performing well with all feature extraction techniques than other classifiers for DS1. Among feature extraction techniques, bigrams are performing well than other techniques.

From Table 6, we observed observe that for DS1, the random forest classifier gives the best result on using bigrams. While in the case of DS2, it gives better result

Table 6 Performance metrics: ensemble classifiers

| Classifier | Feature | A | P | R | F | |
|------------|----------|--------------|-------|-------|-------|-------|
| DS1 | RFC | Bag of words | 0.842 | 0.855 | 0.827 | 0.841 |
| | | TF-IDF | 0.840 | 0.856 | 0.828 | 0.842 |
| | | Bigrams | 0.854 | 0.868 | 0.837 | 0.853 |
| | Bagging | Bag of words | 0.793 | 0.801 | 0.784 | 0.793 |
| | | TF-IDF | 0.795 | 0.802 | 0.788 | 0.795 |
| | | Bigrams | 0.808 | 0.816 | 0.800 | 0.808 |
| | AdaBoost | Bag of words | 0.826 | 0.818 | 0.843 | 0.830 |
| | | TF-IDF | 0.825 | 0.820 | 0.842 | 0.830 |
| | | Bigrams | 0.837 | 0.830 | 0.852 | 0.841 |
| DS2 | RFC | Bag of words | 0.910 | 0.977 | 0.924 | 0.950 |
| | | TF-IDF | 0.901 | 0.962 | 0.927 | 0.953 |
| | | Bigrams | 0.861 | 0.964 | 0.881 | 0.921 |
| | Bagging | Bag of words | 0.772 | 0.948 | 0.795 | 0.865 |
| | | TF-IDF | 0.782 | 0.938 | 0.817 | 0.873 |
| | | Bigrams | 0.772 | 0.948 | 0.796 | 0.872 |
| | AdaBoost | Bag of words | 0.728 | 0.938 | 0.817 | 0.873 |
| | | TF-IDF | 0.891 | 0.966 | 0.924 | 0.925 |
| | | Bigrams | 0.821 | 0.941 | 0.860 | 0.898 |

on using a bag of words. Also, we can see that for DS1, the bagging classifier gives the best result on using bigrams. While in the case of DS2, it gives better result on using bigrams. We can see that for DS1, the AdaBoost classifier gives the best result on using bigrams. While in the case of DS2, it gives better result on using TF-IDF.

6 Conclusion

The objective of this paper is to find the best classifier to test the reviews of movies given out by people so that we would know the overall general opinion of the audience. Out of the two data sets used for experiment, one is already balanced while other is collected manually. All the classifiers are validated using the same metrics. For the first data set, NuSVC classifier outperforms all other classifiers in the set of individual classifiers. While in the case on ensemble classifiers, the random forest classifier outperforms the remaining ones.

When compared between individual and ensemble classifiers, the individual classifiers perform better. For the second data set, SVC classifier outperforms all other classifiers in the set of individual classifiers. While in the case on ensemble classifiers, the random forest classifier outperforms the remaining ones. When compared

between individual and ensemble classifiers, the individual classifiers perform better with a very close difference. We have considered only two data sets. Data sets of different sizes need to be studied for better results.

Better features can be extracted, which are less in number, yet give acceptable results. The set of classifiers can be used collaboratively to get effective results. Changes can be made from the very algorithmic level of the classifiers to gain better performance in the domain of study.

References

1. Sisodia, D. S., & Reddy, R. (2019). Analysis of public sentiments about mega online sale using tweets on big billions day sale. In *Sentiment Analysis and Knowledge Discovery in Contemporary Business* (pp. 59–76). IGI Global.
2. Pak, A., & Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. *LREc*, 10(2010), 1320–1326.
3. Jiang, W. (2014). Study on identification of subjective sentences in product reviews based on weekly supervised topic model. *Journal of Software*, 9(7), 1952–1959.
4. Bhoir, P., & Kolte, S. (2015). Sentiment analysis of movie reviews using lexicon approach. In *2015 IEEE International Conference on Computational Intelligence and Computing Research (ICCIIC)* (pp. 1–6). IEEE.
5. Singh, V., Piryani, R., Uddin, A., Waila, P. (2013). Sentiment analysis of movie reviews and blog posts. In *2013 IEEE 3rd International Advance Computing Conference (IACC)* (pp. 893–898). IEEE.
6. Trupthi, M., Pabboju, S., & Narasimha, G. (2016). Improved feature extraction and classification sentiment analysis. In *2016 International Conference on Advances in Human Machine Interaction (HMI)* (pp. 1–6). IEEE.
7. Sisodia, D. S., & Reddy, N. R. (2017). Sentiment analysis of prospective buyers of mega online sale using tweets. In *2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI)* (pp. 2734–2739). IEEE.
8. Riloff, E., & Wiebe, J. (2003). Learning extraction patterns for subjective expressions. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing* (pp. 105–112). Association for Computational Linguistics.
9. Yu, H., & Hatzivassiloglou, V. (2003). Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing* (pp. 129–136). Association for Computational Linguistics.
10. Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing* (Vol. 10, pp. 79–86). Association for Computational Linguistics.
11. Ghorbel, H., & Jacot, D. (2011) Sentiment analysis of french movie reviews. In *Advances in Distributed Agent-Based Retrieval Tools* (pp. 97–108). Springer.
12. Leung, T., & Malik, J. (2001). Representing and recognizing the visual appearance of materials using three-dimensional textons. *International Journal of Computer Vision*, 43(1), 29–44.
13. Thompson, V. U., Panchev, C., & Oakes, M. (2015). Performance evaluation of similarity measures on similar and dissimilar text retrieval. In *2015 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K)* (Vol. 1, pp. 577–584). IEEE.

14. McCallum, A., Nigam, K., et al. (1998). A comparison of event models for Naive Bayes text classification. In *AAAI-98 Workshop on Learning for Text Categorization* (Vol. 752, pp. 41–48). Madison, WI.
15. Kibriya, A.M., Frank, E., Pfahringer, B., & Holmes, G. (2004). Multinomial Naive Bayes for text categorization revisited. In *Australian Conference on Artificial Intelligence* (Vol. 3339, pp. 488–499). Springer.
16. Wang, X., & Tian, J. (2012). A gene selection method for cancer classification. *Computational and Mathematical Methods in Medicine*, 2012.
17. Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, 1(1), 81–106.
18. Michel, P., & El Kaliouby, R. (2003) Real time facial expression recognition in video using support vector machines. In *Proceedings of the 5th International Conference on Multimodal Interfaces* (pp. 258–264). ACM.
19. Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In *Machine Learning: ECML-98* (pp. 137–142).
20. Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
21. Sisodia, D. S., Vishwakarma, S., & Pujahari, A. (2017). Evaluation of machine learning models for employee churn prediction. In *2017 International Conference on Inventive Computing and Informatics (ICICI)* (pp. 1016–1020). IEEE.
22. Kaggle Inc. Retrived in 2017, <https://www.kaggle.com/c/sentiment-analysis-on-movie-reviews/data>.
23. Mouthshut movie reviews data set. Retrived in 2017, <https://www.mouthshut.com/hindi-movies>.

A Comparative Study on Different Approaches of Road Traffic Optimization Based on Big Data Analytics



Tapajyoti Deb, Niti Vishwas and Ashim Saha

Abstract The emergence of big data has led to technological advancements in various fields including transportation systems. Traffic congestion is a noteworthy issue in numerous urban communities of India alongside other nations. Improper traffic signals usage, poor law enforcement, and poor traffic administration cause movement blockage. Elevated amounts of activity increment stress bring down the quality of life and make a city less engaging. Traffic engineers are accused of influencing transportation frameworks to keep running as effectively as could be expected under the circumstances; however, the assignment appears to be unmanageable. The interconnected technologies around the digital devices offer potential to optimize the road traffic flow. So, there is a need to develop systems for smarter living experience. In order to optimize traffic flow, the first step is to identify the vehicles and then count the traffic at particular intervals so that in case of a jam, the commuters should know the traffic situation and be able to take an alternate route in advance. In this research work, a comparative study on different approaches initiated for road traffic optimization is undertaken along with their advantages and drawbacks which will be beneficial in developing and improving real-time traffic system in near future. It is concluded that the big data analytics architecture for the acquisition and monitoring of real-time traffic information provides the ability to integrate various technologies with existing communications infrastructures, which can help reduce casualties, minimize congestion, and increase safety across street networks capacity and adequacy.

Keywords Big data · Big data analytics · Transportation system · Real-time traffic systems · Traffic flow optimization

T. Deb (✉) · N. Vishwas · A. Saha
National Institute of Technology, Agartala, India
e-mail: tapajyotideb@gmail.com

N. Vishwas
e-mail: niti.vishwas@gmail.com

A. Saha
e-mail: ashim.cse@nita.ac.in

© Springer Nature Singapore Pte Ltd. 2020
M. Pant et al. (eds.), *Performance Management of Integrated Systems and its Applications in Software Engineering*, Asset Analytics,
https://doi.org/10.1007/978-981-13-8253-6_11

1 Introduction

The general population likes to queue, it's said. However, this tolerant attitude to stand in line does not reach everyone who hates a jam to sit in traffic. Congestion is the enemy for city organizers, tenants, and travelers around the world. Traffic congestion is a noteworthy issue in numerous urban communities of India alongside other nations. Improper traffic signals usage, poor law enforcement, and awful traffic administration cause movement blockage. Congestion is bad for business.

What's more, it is additionally destructive to urban flexibility, affecting contrarily on both ecological and social manageability, in terms of emissions and global warming, in addition to air quality and general well being. With respect to the livability of a current city, traffic jam is successfully part of the urban transport user experience. Elevated amounts of activity increment stress bring down the quality of life and make a city less engaging.

Traffic engineers are accused of influencing transportation frameworks to keep running as effectively as could be expected under the circumstances; however, the assignment appears to be unmanageable. In any case, with mechanical advancements and complex data analytics, governments are better ready to fight gridlock now than any time before.

In the long run, big data analytics will make driving considerably more secure. Thus, ample opportunity is there to adequately deal with the traffic blockage issue. Prescient analytics provide a completely new way to extract valuable data on urban mobility from networked infrastructure, associated vehicles, and cell phones, to evaluate traffic patterns in real time and to implement the necessary management strategies. Data sources in our cities mean that analytics can open up an entirely new era of intelligent transport.

2 Importance of Big Data Analytics in Road Traffic Scenario

Managing traffic in a smart way using big data analytics becomes essential as the conventional signaling system is not effective in the heavy traffic roads. A smart traffic management system is helpful not only to reduce traffic congestion and also to ensure the smooth flow of traffic at peak times and during emergency situations, such as following an accident, for instance.

The analysis information and predictions from these systems can be projected in real time on digital screens installed at city center entrances or on mobile devices, leading drivers to available parking lots and streets. This will not only reduce congestion, but also save time and fuel, making the environment cleaner and better to live. Therefore, an intelligent traffic system will increase the ease of living.

3 Literature Survey

3.1 International Status

Urban areas have a few difficulties for us to handle; one of them is the issue of urban portability. The United Nations anticipated that half of the total populace would live in urban zones toward the finish of 2008. The developing populace and the absence of accessible physical space have made traffic administration progressively difficult in Singapore. By 2020, travel demand is supposed to ascend from 8.9 million trips for each day to around 14.3 million, implying the stamped increment in the city–state’s populace.

Simultaneously, Singapore confronts real imperatives in space, with 12% of land officially possessed by the 3300-km street arrange and another 15% dedicated to lodging [1]. Extending the street system to address transport request has not been viewed as an economical alternative. Rather, the Singapore government has used approach and innovation to oversee transport request and supply, amplifying the current manageable frameworks while limiting the more environmentally impactful methods of travel. These days, forecasts say that by 2050, 86% of the developed world and 64% of the developing world will be urbanized. The figure beneath demonstrates the growth rate of urban and rural population [2].

Diverse urban areas and districts are additionally working together to fabricate interconnected savvy traffic administration systems that improve the general street limit inside a region and between locales. For instance, the city of Amsterdam now utilizes traffic link’s SCM framework that is associated with the traffic activity of the national government. Both centers can see and consequently manage traffic within the region. The city intends to grow the support of in-auto and navigation hardware to additionally lessen blockage.

Yarra Trams, Administrator of Melbourne, Australia’s 100-year-old tram network, is utilizing Internet of things (IOT) innovation to diminish traffic congestion and improve passenger experience. It is the biggest working tram network in the world with more than 250 km of twofold tracks. It oversees more than 91,000 bits of hardware and 487 trams going on 29 distinct courses.

3.2 National Status

The shortage of traffic cops and an increasing number of vehicles in the city lead to heavy traffic jams during peak hours every day at all major crossings. The Municipal Corporation of Gurgaon (MCG) will partner with a Japanese company to install an integrated intelligent traffic system (IITS) at various crossings in order to avoid such shocking situations (Fig. 1).

Urban and rural population of the world, 1950–2050

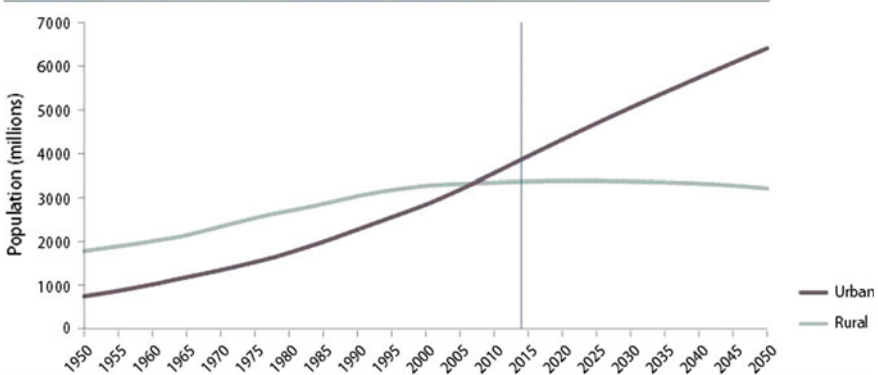


Fig. 1 Estimated growth of population of the world in urban and rural areas

The system intends to control traffic according to traffic volume by warning drivers and changing signals. The routes have induction loops that are rope coils embedded under the road, so the counter increases every time a vehicle passes.

Police in Delhi have introduced a number of technological innovations such as the online prosecution system, the traffic control system, and the information dissemination SMS facility and the accident information system based on GIS.

The state government of Bangalore has promised \$15.8 million to improve traffic management in Bangalore. There will be around 100 enforcement cameras and 800 solar traffic signals operated by the vehicle. And collected traffic information is shared on real-time via SMS and FM radio. India's freight volume is increasing annually at a rate of 9.08%, according to India's transport company and IIM and that of vehicles at 10.76%, but that of the road is only by 4.01%. This has led to a reduction in road space according to the number of vehicles in total. India's average fuel mileage is just 3.96 km/l. The main reason for traffic congestion is that India is the second most populated country in Asia after China, so the number of vehicles is also increasing with an increase in population.

Different Approaches of Road Traffic Optimization Based on Big Data Analytics Mining GPS Data for Traffic Congestion Detection and Prediction, 2013, By Suhas Prakash Kaklij

The traffic congestion and detection model proposed by them are structured and applied over GPS data. The data being collected are first categorized using k-means clustering algorithm, and the obtained clusters are filtered out. Further, Naive Bayes algorithm is used as mining method to detect and predict traffic congestion [3].

The steps of the algorithm as proposed by authors are as follows:

1. The initial GPS data are collected from devices like mobile phones, on-board units in the form of log files with information of device id, location details like latitude, longitude, time, speed, and day.

2. The attributes are then stored to the database and haversine formula is used to calculate distance between two points.

$$\text{hav}(d/r) = \text{hav}(\varphi_2 - \varphi_1) + \cos(\varphi_1) \cos(\varphi_2) \text{hav}(\lambda_2 - \lambda_1) \quad (1.1)$$

where hav is the haversine function.

3. The data are then inputted to the filtering process where k-means algorithm is applied to get clusters.
4. Finally, Naive Bayes algorithm is applied to these clusters for traffic congestion detection.

The framework proposed identifies constant traffic congestion locations adopting the data coming in from various types of GPS-enabled devices. Various modules are used to isolate on-road and off-road traffic as well. The framework is flexible to different cities by changing the city-dependent thresholds.

Flaw(s): The framework lacks the use of Hadoop architecture and MapReduce functions which can drastically reduce the computation time of the log data. No proper methodologies are proposed for traffic prediction.

Scalable Traffic Video Analytics Using Hadoop MapReduce, 2015, By Vaithilingam Anantha Natarajan et al.

This paper illustrates video analysis of road traffic using Hadoop. The data include traffic volume, traffic speed, and speed of each vehicle. The system detects road congestion and suggests alternate route to relevant commuters. The system is deployed on a cluster computer [4].

The steps of the algorithm as proposed by authors are as follows:

1. Initially, the video is captured from the IP cameras across traffic points and then the video streams are sent to the media server and finally to the supervising station.
2. The video streams are splitted and put through MapReducer which outputs small video chunks that are written to HDFS.
3. In the vehicle detection phase, type of vehicle, speed, and congestion are detected. Vehicle detection is done using HAAR classifier [5] and SVM. Speed of vehicle is detected from the displacement of centroids, and overlapping of two centroids below a threshold speed guarantees a collision.
4. The traffic statistics data are then sent to the hive tables. In case of congestion, alternate routes are computed from these data and sent to the commuters.

Benefit(s): The suggested work demonstrated live video analysis of road traffic video for collision detection and congestion avoidance. The accuracy of the detection process was found to be satisfactory, and the usage of multiple nodes reduced the processing time of the data.

Flaw(s): Addition of few more nodes would have resulted in faster execution time.

Big Data Approach for Secure Traffic Data Analytics, 2016, By Koushalya Bijjaragi and Poonam Tijare

This paper deals with traffic analysis and giving useful predictions to end user [6]. Hadoop Distributed File System (HDFS) is used to store the data, and MapReduce [7] framework is used for parallel processing.

The steps of the algorithm as proposed by authors are as follows:

1. A user interface is created in the data server where records can be inserted and datasets can be uploaded as well.
2. User signature is verified using a customized algorithm after which the busy/idle traffic algorithm is implemented.
3. The output of the prediction analysis gives the best, average, and worst time to travel on hourly basis traffic density.

Benefit(s): The proposed methodology presented a secure approach to analyze traffic data. Each user is authenticated through signature verification. Big data technologies like Hadoop and MapReduce overcome the existing limitations of traffic administration done manually.

Flaw(s): Data to be operated are not real time. So, the accuracy of predictions will not be satisfactory. Also during traffic congestion, alternative route suggestion will not be feasible since data collected are not in real time.

Big Data Analytics Architecture for Real-Time Traffic Control, 2017, By Sasan Amini, Ilias Gerostathopoulos and Christian Prehofer

This research project proposes a comprehensive and flexible architecture for real-time traffic control based on a distributed computer platform. The architecture is based on a systematic analysis of the needs of existing traffic control systems using Kafka, a large data processing tool and data pipelines [8]. They demonstrated their approach on a case study of controlling freeway hard shoulder lane in simulation [9].

The steps of the algorithm as proposed by authors are as follows:

1. Kafka, a big data tool, is used to decouple intelligent transport system (ITS) into two categories.
2. The data items are then sent to Hadoop Distributed File System (HDFS) data warehouse for analysis.
3. The engine gets input and performs analysis. The results of the analysis are stored in a no SQL database.

The proposed framework provides a comprehensive and flexible architecture based on big data analytics for real-time traffic control. The generated data were well structured, and no quality checks were required. The framework can be expanded to handle additional loads from multiple data sources.

Flaw(s): The main limitation of their work was implementing it on real world which requires analyzing huge datasets from large number of sources. Everything was done in simulation.

Development of Road Traffic Analysis Platform Using Big Data, 2017, By Sung et al.

This paper deals with the prediction of road traffic based on big data analytics. Various big data technologies and their implementation in overseas were discussed. The framework they proposed provides optimized traffic information services different from existing sources. In addition, the platform can use road traffic sensor and unstructured data to provide meticulous road driving environment information to improve the driver's safety and to boost the reliability of traffic prediction information [10, 11].

The steps of the algorithm as proposed by authors are as follows:

1. Vehicles' sensor data were designed to register data in restful way using REST server and Kafka.
2. Public data were designed to be collected and optimized by Flume-based data collection agent.
3. The collected big data is sent to Hadoop Distributed File System (HDFS) or no SQL for analysis.
4. Visualization of analysis and display of information are provided using Web-based GIS map.

Benefit(s): The suggested prediction platform can provide flexible traffic information services for real-time traffic control, thereby reducing congestion. The platform can use road traffic sensor information and related unstructured data to provide information which can boost the reliability of prediction.

Flaw(s): The prediction model lacks implementation in customized environment which requires analyzing huge datasets' multiple sources. There is also a need to develop more reliable traffic data platforms through a continuous verification process.

4 Conclusion

Big data analysis for the acquisition and monitoring of real-time traffic information architecture provides the ability to integrate various technologies with existing communication infrastructures, which can help reduce casualties, minimize congestion, and increase the security of street networks and their capacity and adequacy. This paper shows different methods of implementing road traffic optimization techniques, and also we get to know the hindrances and drawbacks that may arise.

Apache Hadoop and MapReduce framework will help in distributed processing, storing, and sorting the outputs of the maps, which are then inputted to the reduce tasks. These technologies along with the other's work will be beneficial for further research and development of intelligent traffic system which will help diminish traffic jam and casualties making driving across road much faster and safer.

References

1. Intelligent Transportation Systems Next Wave of the Connected Car. <https://www.intellimec.com/insights/intelligent-transportation-systemsnext-wave-connected-car/>.
2. Urbanization of World <http://www.civilservicesexpert.com/geography/urbanisation-of-world/>.
3. Kaklij, S. P. (2015). Mining GPS data for traffic congestion detection and prediction. *International Journal of Science and Research (IJSR)*, 4(9), ISSN:2319-7064. (Online).
4. Natarajan, V. A., Jothilakshmi, S., & Gudivada, V. N. (2015). Scalable traffic video analytics using Hadoop MapReduce. *ALLDATA 2015: The First International Conference on Big Data, Small Data, Linked Data and Open Data*.
5. Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Vol. 1, pp. I-511–I-518).
6. Bijjaragi, K., & Tijare, P. (2016) Big data approach for secure traffic data analytics using Hadoop. *International Journal on Recent and Innovation Trends in Computing and Communication*, 4(5). ISSN 2321-8169.
7. Dean, J., & Ghemawat, S. (2008). MapReduce: Simplified data processing on large clusters. *Commun. ACM*, 51(1), 107–113.
8. Kreps, J., Narkhede, N., & Rao, J., et al. (2011) Kafka: A distributed messaging system for log processing. In *Proceedings of the NetDB* (pp. 17).
9. Amini, S., Gerostathopoulos, I., & Prehofer, C. (2017). Big data analytics for real- time traffic control. In *5th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS)*. <https://doi.org/10.1109/mtits.2017.8005605>.
10. Sung, H. K., & Chong, K. S. (2017). Development of road traffic analysis platform using big data. In *International Conference on Advances in Big Data Analytics* (pp. 65–66). CSREA Press ©, ISBN: 1-60132-448-0.
11. Kim, R., & Kang, M. M. (2014). Today and the future of big data analytics technology. *The Korean Institute of Information Scientists and Engineering*, 1, 8–17.

Comparative Study Between Cryptographic and Hybrid Techniques for Implementation of Security in Cloud Computing



Sumit Chaudhary, Foram Suthar and N. K. Joshi

Abstract Cloud computing depicts the latest technology in the field of information communication technology. Cloud computing technology offers the user with an endless list of services which they can avail, and these services range from hardware and software and infrastructure to resource utilization. Cloud computing allows us to access online software application, data storage, and processing power of system from anywhere and at anytime. Cloud computing supports the organization to increase their capacity remotely without creating own infrastructure, platform, purchasing new licensed software that is required for automation of various processes. There are a number of dominant parameters like energy dissipation, resource allocation, virtualization, and security that a user needs to keep in mind while selecting cloud computing services. Of the dominant parameters, security as a parameter in cloud holds a special concern and is one of the biggest challenges in cloud computing as far as data transfer process and data storage process are a concern as it happens through the Internet. In this research work, the authors have presented a comparative study between cryptographic and hybrid techniques for security concerns of the cloud computing paradigm. Through the results obtained from the comparative study of these techniques, it is observed that the effectiveness of the hybrid techniques is better in terms of execution of the algorithm than the cryptographic techniques (common public- and private-key cryptography) of security implementation in cloud computing.

Keywords Cloud computing · Security · Cryptography · Cryptographic security techniques · Hybrid security techniques

S. Chaudhary (✉) · N. K. Joshi
Uttaranchal University, Dehradun, India
e-mail: iimtsmit@gmail.com

N. K. Joshi
e-mail: nkjoshi2001@yahoo.com

F. Suthar
Indrashil Institute of Science & Technology, Cadila Group, Ahmedabad, India
e-mail: foram.suthar@iist.edu.in

1 Introduction

Presently, cloud computing is the latest trend in IT sector. The Internet has been graphically derived by cloud symbol since many years. Cloud shows that the data are transferred over the Internet in a computer network. Computing is processed to utilize computer network to perform our task, and it can be done by hardware device–software device. Computing is done by people every day in the life like to perform calculation automatically, transfer a data, sending a mail, swapping credit card, etc. Cloud computing allows us to access online software application, data storage, and processing power of system at anytime and anywhere. Thus, the name cloud computing is inspired by the symbol of the cloud where the computing will happen over the Internet.

The US National Institute of Standards and Technology (NIST) [1] defines cloud computing as follows: ‘cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction’ [2]. Cloud computing supports the organization to increase their capacity remotely without creating own infrastructure, platform, purchasing new licensed software that is required for automation of various processes [3].

Cloud is a platform where the user can store the data and use the resources virtually. Cloud allows the user to pay for whatever resources they use, and the user can expand and scale down the resources according to needs. Cloud allows client to access different services using different service delivery model: infrastructure as a service (IaaS), platform as a service (PaaS), software as a service (SaaS) [4], and deployment model—private cloud, public cloud, and hybrid cloud [5].

Cloud support several essential characteristics like on-demand self-service, broad network access, rapid elasticity, resource pooling, and measured services to the user. One of the biggest challenges in cloud computing is data security because data transfer process and data storage process are happening through the Internet. So, it is difficult to secure our data and resource against an unauthorized person, masquerade, and eavesdropper. There are several security issues that occur in cloud computing like physical security lost because we do not know where the resource runs. Data integrity is lost because common standard to ensure data integrity does not exist [6]. There are several security algorithm available to prevent the data security while accessing Web in cloud computing. Detail description of preventing techniques is mentioned below.

2 Literature Review

Dimitrios et al. study cloud security in depth in 2012 [7]. The authors proposed security solution by using trusted third party. The solution is based on single sign-on

(SSO) mechanisms and the lightweight directory access protocol (LDAP) [8] that is cryptography specifically in public-key infrastructure. This is to make sure the security proposed solution consists of using cryptography to ensure confidentiality and integrity of involved data. However, it does not recognize which encryption algorithms used. But, this solution does not recognize the encryption algorithm to be used.

Cunsolo et al. in 2009 [9] came up with mechanism that protected data in distributed systems such as grid, cloud, and autonomic. In this paper, author implements the techniques using symmetric and asymmetric algorithm. The limitation of this technique is that the concept of sharing resources is contradicted because data access was done by owner only in the cloud environment.

Hashizume et al. [10] classified different cloud service models (SaaS, PaaS, and IaaS) to solve security issues. The relation between cloud layers and the common threats shows the main vulnerabilities in cloud computing. The solution of splitting some available countermeasures is a technical implementation, which is not covered in this study.

Rahmani et al. [11] implemented a new technique, encryption as a service (EaaS), as a solution based on XaaS concept for cryptography in cloud computing. The security risks and the inefficiency of cloud provider's encryption and of client-side encryption can be prevented by this solution, respectively. Moreover, this solution does not show a comparative study of cryptographic algorithms that can be integrated.

The performances of cryptographic algorithms in cloud platform are evaluated using symmetric and asymmetric algorithms by the Mohammad et al. [12]. Different encryption techniques which are based on key size, the performance, and the size of the output file are discussed in this paper. The distribution of encryption keys in a secure way is not proposed, but it proposed AES algorithm to encrypt data for more security.

In [7] paper, secure cloud architecture using cryptography was first proposed by this D. Zissis. For cloud storage, they proposed to use cryptographic algorithms [11, 13]. But, they do not specify which algorithm is recommended to encrypt data and how to distribute cryptographic keys while maintaining adequacy with cloud characteristics. So these solutions remain incomplete.

In [2], Belguith proposed a new lightweight cryptographic algorithm which is the combination of AES as public-key algorithm to encrypt data and RSA as public-private-key algorithm to distribute keys. During conserving the rights of users to access data by a secured and authorized way, this combination helps to benefit from the efficient security of asymmetric encryption and the rapid performance of symmetric encryption.

3 Cryptographic Techniques

(a) Data Encryption Standard (DES)

DES is proposed by NIST in 1977. DES is block cipher and follows 16-round Feistel cipher structure. DES is based on symmetric key where the same key is used for encryption and decryption. This algorithm support 64-bit plaintext and 56-bit key value. First 64-bit plaintext value goes through the initial permutation (IP) stage which reorders the bits and generates permuted input. After initial permutation your 64-bits input value divided into two halves of 32-bits [14].

In key generation process, the function will ignore every 8-bit from original 64-bit key value and generate 56-bit key from 64-bit key. Those 8-bit key value is further used as parity bits in DES algorithm; 56-bit key value is further divided into two halves of 28 bits, and individual value will rotate 1 or 2 bits by circular left shift register. In next step, 56-bit key value passes to the permutation that produces final 48-bit key. Now, right halves of 32-bit plaintext value pass through expansion permutation where you bits will expand from 32 bits to 48 bits. These 48-bit plaintext and 48-bit key value will be XORed and substituted into 32-bit plaintext by S-box.

Final 32-bit plaintext value passes from permutation P-box and XOR with left halves of 32-bit plaintext. This process will do till 16 rounds, and 48-bit key value will be changed at each round. The heart of this algorithm is an inner function which includes expansion, substitution, and permutation process.

DES is less secure algorithm because it supports 56-bits key that is too short. It is vulnerable to brute-force attack.

(b) Triple DES (3DES)

Triple DES is similar to DES, but it includes three stages of encryption and decryption with two different keys k_1 and k_2 . First, plaintext will be encrypted by k_1 key; then, plaintext will be decrypted by k_2 key, and at last, it will gain encrypted by k_1 key. Triple DES overcomes the security problems of DES. To crack DES through brute-force attacks, 2^{112} operation is required, so it is still be secure.

(c) Advanced Encryption Standard (AES)

AES is developed by two people Joan Daeman and Vincent Rijmen in 2000 which is also called Rijndael cipher. AES is extended of DES and published by NIST in 2001 [15]. It is block cipher and supports 128-bit block as plaintext. Except DES, AES supports three key sizes—128, 192, or 256 bits. It is a symmetric key cryptography, which supports both confusion and diffusion. On base of key value, it works on three rounds—10, 12, or 14. In symmetric key cryptography (AES), 128-bit plaintext passes through different four stages in one round as shown below.

1. **Substitution bytes:** AES uses 16×16 matrix of byte values, called as S-box to perform a byte-by-byte substitution [16]. This S-box contains permutation of all possible 256 8-bit values. 128-bit plaintext value is divided into 16 blocks of 8 bits. The leftmost 4 bits are considered as row value, and

- rightmost 4 bits are considered as column value [17]. Unique 8 bit output value generated by row and column value which are indexes into the S-Box.
2. **ShiftRows:** ShiftRows process is simple permutation process. In ShiftRows process, 4×4 matrixes are generated using unique 128 bits which are generated by substitution byte process. This stage performs process as its name shows. First row of state will remain same. The value of second row will shift left by 1 byte. For the third row, 2-byte left shift is performed. And four rows shifted left by 3 bytes.
 3. **MixCloumns:** MixCloumns process is simple substitution process. Final 128 bits which are generated by ShiftRows stage pass through MixCloumns process. Each byte of column is multiplied with new function value.
 4. **AddRoundKey:** In AddRoundKey bits value XOR with key value (128, 192, 256 bits).

These four stage processes are repeated till the number of rounds 10, 12, and 14 depends on key values of 128, 192, and 256 bits. AES is completely secure against the brute-force attacks because it is a required multibit key to encrypt the data.

(a) **Rivest–Shamir–Adleman Algorithm (RSA)**

Three researchers Rivest, Shamir, and Adleman proposed secure public-key cryptography algorithm. RSA is block cipher where plaintext and ciphertext are integers between $[0, n - 1]$ for some n where size of n is 1024 bits. RSA supports public–private-key (asymmetric cryptography algorithm) concept where encryption will be done by public-key and decryption will be done by private-key. Key generation schema of RSA is shown below.

1. Choose two very large digit prime numbers a and b , where $a \neq b$
2. Calculate: $x = a*b$.
3. Calculate: $\phi(x) = (a-1)*(b-1)$
4. Select integer p Where, $\gcd(\phi(x), p) = 1$; $1 < p < \phi(x)$
5. Calculate n where, $q = p-1(\text{mod } \phi(x))$
6. Now, we have Public Key $PU = \{p, x\}$ Private Key $PR = \{q, x\}$
7. Encrypt plaintext M and create cipher text: $C = M*p \text{ mod } x$
8. Decrypt cipher text C and create plain text $M = C*q \text{ mod } x$

RSA algorithm is very popular and secure because it is very difficult to factor very large numbers. Factoring large number is not a difficult task, but today not a single algorithm exists to factor a 200-digit number in the reasonable amount of time. This algorithm can be applied to any electronic fund transmissions [18].

(b) **Diffie–Hellman Key Exchange (D-H)**

Diffie and Hellman published a new public-key algorithm which enables two users to establish secret key [19]. This is one of the first secret key transfer protocol over a public channel. When two parties want to share their data over network, first they generate secret key and transfer that secret key in secure channel. Data transfer process will happen between two users when both parties

get keys securely by D-H algorithm [20]. Key generation schema between two person M and N is shown below.

1. Choose two random number: a, b where a is prime number and b is primitive roots of a and $b < a$
2. Select private key for user M: X_M where $X_M < a$
3. Calculate public key for user M: Y_M where $Y_M = b^{X_M} \text{ mod } a$
4. Select private key for user N: X_N where $X_N < a$
5. Calculate public key for user N: Y_N where $Y_N = b^{X_N} \text{ mod } a$
6. Compute secret key by user M: $K = (Y_N)^{X_M} \text{ mod } a$
7. Compute secret key by user N: $K = (Y_M)^{X_N} \text{ mod } a$

In DH key exchange protocol secure because to calculate exponentials modulo of a prime is easy task, but calculation of discrete logarithms is very difficult.

4 Hybrid Techniques

The basic idea behind hybrid techniques is improving the efficiency of the existing algorithm. Many hybrid techniques are used to encrypt the data; some of them are explained below.

(a) Dual RSA

It is a new public-key cryptography algorithm; it is also called RSA-CRT, because it uses Chinese remainder theorem, CRT, for its decryption [21]. Dual RSA has been developed for better performance in terms of computation costs and memory storage requirements. RSA takes one block at a time to encrypt and decrypt the data. But dual RSA takes two blocks during encryption and decryption [21]. Encryption time of dual RSA is more as compared to decryption of two blocks. Thus, dual RSA increases the performance as compared to RSA.

(b) AES-RSA

Symmetric algorithms are faster to encrypt data as compared to asymmetric techniques. RSA (asymmetric algorithm) takes more processing time to encrypt data due to longest key size. A new symmetric and asymmetric hybrid model has been developed to increase the level of security. In AES-RSA [22] hybrid techniques, AEs first generate 256-bit key. That 256-bit key is expanded to 1024-bit key and is used in RSA as private-key.

(c) RSA-AES-Digital Signature

Digital signature [23] is an authentication mechanism which provides data integrity and authentication. It is public-private-key algorithm where data are encrypted by sender's public-key and decrypted by receiver's private-key. In this hybrid key generation algorithm, first 1024-bit key is generated by RSA and XOR with 1024-bit key of AES. Final hybrid key generated by RSA and AES algorithms is used in digital signature as sender's private-key (Table 1; Fig. 1).

Table 1 Efficiency of cryptographic algorithm

| Algorithm | Efficiency |
|------------|------------|
| DES | 45 |
| 3DES | 20 |
| AES | 65 |
| RSA | 10 |
| Dual RSA | 30 |
| AES-RSA | 78 |
| RSA-AES-DS | 80 |

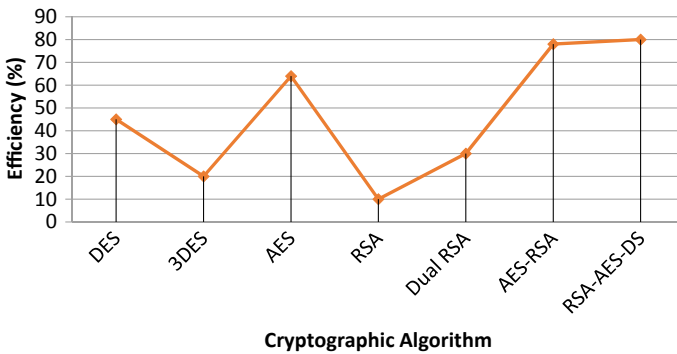


Fig. 1 Comparison of cryptographic and hybrid cryptographic techniques

5 Conclusion and Future Scope

Data security is major issue in cloud computing. Many techniques are available to make secure communication, but one of the most efficient techniques is hybrid technique. This paper explores the comparison between cryptography and hybrid cryptography techniques. Based on the result, we can conclude that the efficiency of symmetric algorithm is more as compared to asymmetric algorithm, and the efficiency of hybrid technique gives average efficiency in any procedure. Because hybrid technique is a combination of public-key cryptography and private-key cryptography, then the security and avalanche effect will be more and also with the use of hybrid technique non-linear function generated.

The work is defined in this paper in terms of efficiency, but we can extend the same work in the different parameters like security, power saving, scheduling, and complexity. Also, we can show the same work with any simulation tool for showing all the parameters, and we can analyze each hybrid technique for different parameters.

References

1. Mell, P., & Grance, T. (2011). *The NIST definition of cloud computing*. Information Technology Laboratory: National Institute of Standards and Technology.
2. Belguith, S., Abderrazak, J., & Attia, R. (2015). Enhancing data security in cloud computing using a lightweight cryptographic algorithm. In *The Eleventh International Conference on Autonomic and Systems*.
3. Pai, T., & Aithal, P. S. (2017). Cloud computing security issues-challenges and opportunities.
4. Kanday, R. (2012). A survey on cloud computing security. In *2012 International Conference on Computing Sciences (ICCS)*. IEEE.
5. Sudha, M., & Monica, M. (2012). Enhanced security framework to ensure data security in cloud computing using cryptography. *Advances in Computer Science and its Applications*, 1(1), 32–37.
6. Arora, R., Parashar, A., & Cloud Computing Is Transforming. (2013). Secure user data in cloud computing using encryption algorithms. *International Journal of Engineering Research and Applications*, 3(4), 1922–1926.
7. Zissis, D., & Lekkas, D. (2012). Addressing cloud computing security issues. *Future Generation Computer Systems*, 28(3), 583–592.
8. “Directories and Public–Key Infrastructure (PKI)”, VeriSign, 2004.
9. Cunsolo, V. D., Distefano, S., Puliafito, A., & Scarpa, M. (2009). Achieving information security in network computing systems (pp. 71–77). In *Eighth IEEE International Conference on Dependable, Autonomic and Secure Computing (DASC'09)*.
10. Hashizume, K., Rosado, D. G., Fernández-Medina, E., & Fernandez, E. B. (2013). An analysis of security issues for cloud computing. *Journal of Internet Services and Applications*, 4, 1–13.
11. Rahmani, H., Sundararajan, E., Ali, Z. M., & Zin, A. M. (2013). Encryption as a Service (EaaS) as a solution for cryptography in cloud. *Procedia Technology*, 11, 1202–1210.
12. Mohammad, J., Omer, K., Abbas, S., El-Horbaty, E. S. M., & Salem, A. B. M. (2013). A comparative study between modern encryption algorithms based on cloud computing environment. In *8th International Conference for Internet Technology and Secured Transactions (ICITST'13)* (pp. 531–535). IEEE.
13. Singh, G. (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.
14. Kumar, G., Rai, M., & Lee, G. (2011). An approach to provide security in wireless sensor network using block mode of Cipher. *Security Technology*, 101–112.
15. Elminaam, D. S. A., Abdual-Kader, H. M., & Hadhoud, M. M. (2010). Evaluating the performance of symmetric encryption algorithms. *IJ Network Security*, 10(3), 216–222.
16. Shivkumar, S., & Umamaheswari, G. (2011) Performance comparison of Advanced Encryption Standard (AES) and AES key dependent S-box-Simulation using MATLAB. In *2011 International Conference on Process Automation, Control and Computing (PACC)*. IEEE.
17. Yifeng, Bai, Xiao Jian, and Yu Long. Kernel partial least-squares regression. In *International Joint Conference on Neural Networks. IJCNN'06*. IEEE.
18. Milanov, E. (2009). *The RSA algorithm*. RSA Laboratories.
19. Rescorla, E. (1999). *Diffie-hellman key agreement method*.
20. Van Der Merwe, J., Dawoud, D., & McDonald, S. (2005). Fully self-organized peer-to-peer key management for mobile ad hoc networks. In *Proceedings of the 4th ACM Workshop on Wireless Security*. ACM.
21. Subasree, S., & Sakthivel, N. K. (2010). Design of a new security protocol using hybrid cryptography algorithms. *IJRRAS*, 2(2), 95–103.
22. Al Hasib, A., & Haque, A. A. M. M. (2008). A comparative study of the performance and security issues of AES and RSA cryptography (Vol. 2). In *Third International Conference on Convergence and Hybrid Information Technology. ICCIT'08*. IEEE.

23. Somani, U., Lakhani, K., & Mundram, M. (2010). Implementing digital signature with RSA encryption algorithm to enhance the data security of cloud in cloud computing. In *2010 1st International Conference on Parallel Distributed and Grid Computing (PDGC)*. IEEE.
24. Yifeng, B., Xiao, J., & Yu, L. (2006). Kernel partial least-squares regression. In *International Joint Conference on Neural Networks*. IJCNN'06. IEEE.
25. Ruhai, W. (2006, November). NIS05-1: Performance Analysis of Advanced Encryption Standard (AES). In *IEEE Globecom*.
26. Singh, A., & Misra, A. (2012, January) Analysis of cryptographically replay attacks and its mitigation mechanism. In *Proceedings of the International Conference on Information Systems Design and Intelligent Applications 2012 (INDIA 2012) held in Visakhapatnam, India*. Berlin/Heidelberg: Springer.
27. Kofuji, S. T. (2013). Performance analysis of encryption algorithms on mobile devices. In *2013 47th International Carnahan Conference on Security Technology (ICCST)*. IEEE.
28. Al-Anzi, F., Al-Enezi, M., & Soni, J. (2016). New proposed Z-transform based encryption algorithm. In *2016 6th International Conference on IT Convergence and Security (ICITCS)*. IEEE.

Functional Module Detection in Gene Regulatory Network Associated with Hepatocellular Carcinoma



Sachin Bhatt, Kalpana Singh and Ravins Dohare

Abstract Hepatocellular carcinoma (HCC) is a common type of liver cancer and has a high mortality rate worldwide. Its prognosis remains poor due to tumor recurrence or tumor progression and diagnosed at advanced stage. Hence, there is a critical need to develop effective biomarker for understanding the HCC mechanism. Although the existing evidence demonstrates the important role of single-gene abnormality, often the genes modularity is ignored. In this research work, the authors aim to find modular structure with potential functional relevance. The authors aim to construct a gene regulatory network of DEGs and perform its topological analysis about its hidden structure. The authors have also detected the modules in constructed network, along with their enrichment, finding pathways they are involved in, and their biological functional analysis. There are three major steps adopted for carrying out this research work. Firstly, the authors filtered differentially expressed genes (DEGs) from gene expression data obtained from GEO database which included ten normal and ten HCC samples. Secondly, they constructed co-expressed gene regulatory network (GRN) of DEGs using Pearson's correlation coefficient, then unraveled the characteristics of GRN, and finally, detected modules from GRN and their functional relevance along with the topological characters of network. The DEGs in normal and HCC were identified using MATLAB, while the network is constructed using Cytoscape. The modules in network have been derived from online software GraphWeb.

Keywords Hepatocellular carcinoma · Module selection · Gene regulatory network · Topological measure

S. Bhatt · K. Singh · R. Dohare (✉)
Centre for Interdisciplinary Research in Basic Sciences, Jamia Millia Islamia, New Delhi, India
e-mail: ravinsdohare@gmail.com

S. Bhatt
e-mail: sachin_bhatt@rocketmail.com

K. Singh
e-mail: contact.kalpanasingh@gmail.com

© Springer Nature Singapore Pte Ltd. 2020
M. Pant et al. (eds.), *Performance Management of Integrated Systems and its Applications in Software Engineering*, Asset Analytics,
https://doi.org/10.1007/978-981-13-8253-6_13

1 Introduction

Liver cancer is the second most leading cause of death associated with cancer. During the year of 2012, around 782,000 new cases occurred and estimated to be account for nearly 746,000 deaths worldwide, which account for 9.1% of the total deaths due to cancer. It is the fifth most prevailing cancer type in men accounting for nearly 554,000 new cases, and ninth in women with around 228,000 new incidence cases. Metastasis or secondary liver cancers spread from one part to other parts or tissues of the body and are usually nonspecific. The early detection for liver cancer is very poor with an overall fraction of mortality upon incidence of 95% [1].

HCC is being the most frequent form of liver cancer. The rate of incidence and mortality is very high in HCC. An estimation of 696,000 deaths in 2008 was because of HCC, with nearly 80% of HCC cases occurring in developing countries. The HCC incidence continues to increase worldwide and varies markedly in different regions. The reasons being for the high incidences of HCC are chronic hepatitis B virus (HBV), hepatitis C virus (HCV) infection, and fungal toxin, aflatoxin B1. More than 8% of population is affected by chronic HBV, and one-fourth of them will develop the tumor which may consequence as HCC [2].

The HCC is dispersed according to geographical region. According to World Health Organization 2012 report, HCC is a great worry of less developed regions where approximately 80% of new cancer cases occurred in 2012 worldwide. The occurrence of HCC is more than twice in developing countries than that in developed countries globally. The maximum cases of HCC incidence are in Eastern Asia, with incidence rates in male of 31.9 per 100,000 populations, followed by African region and the Pacific Islands. Global epidemiology of HCC is directed by prevalence of dominant viral hepatitis in the underlying population and also the age at which it is acquired [3]. HBV transmitted at birth is most common cause in the regions of high occurrence of HCC. It is more common in males than females since HBV, HCV, and consumption of alcohol are more common in males, and cirrhosis due to chronic HBV or hepatitis HCV is the leading cause associated with HCC [4].

Despite modern technology for diagnosis and management, the mean survival of HCC patients remains under 8 months [5]. Surgical methods, like transplantation of liver, are the only curative practice for HCC, while its recurrence has the high probability in patients within 5 years after surgery [6]. The prognosis of HCC patients remains poor because of tumor recurrence or tumor progression, and effective adjuvant therapies remain lacking to date [7]. However, the sensitivity and specificity of a single biomarker are inadequate for the clinical diagnosis or prognosis of HCC or liver metastasis. The tumor marker alpha-fetoprotein is used as a supplement in HCC history progression but not in early disease detection [8]. Therefore, understanding the molecular carcinogenic mechanism of HCC is crucial.

The idea of functional component led to the foundation of module [9]. Module at molecular level can be defined as the collection of genes, protein, or any other product that collectively work in a coordinated fashion to bring out the cellular function. A module is thought to be a separate functional entity. Modules having significant biological function are the central player for the understanding more about complex diseases.

Module determination is one of the emerging topics of network. The concept of functional components is the foundation of systems biology; a functional module is a distinct entity which has separate function than the other modules [9]. Modules having meaningful biological functions are the key player in order to understand more about the living beings by developing new hypothesis or processes that are common in biological system. Large amount of interaction of data are being generated rapidly, and integration of such network data with each other and with different molecular profiles in pursuance of molecules that belong to a common biological function is the recent focus of bioinformatics research [10]. In the year 2000, Lau et al. have already attempted to construct for unfolding gene–gene relationship among the genes which are differentially regulated and identifying gene modules for better understanding of molecular mechanisms [11].

Therefore, many studies used computerized approaches based on datasets covering both human PPI networks and cancer gene expression profiles. Network-based analyses of living cells have been employed to detect modules in order to characterize specific physiological functions, signaling and metabolic networks, and genes with clinical significance [12].

Thus, we are constructing a gene regulatory network of DEGs and perform its topological analysis about its hidden structure. The scale-freeness [13] and hierarchical structure are imparted global characteristics to understand the behavior of gene regulatory network [14]. Furthermore, we also detecting the modules in constructed network, along with their enrichment, finding pathways they are involved in, and their biological functional analysis.

2 Materials and Methods

2.1 Gene Expression Data

The gene expression data of HCC were retrieved from Gene Expression Omnibus (GEO), an open genomic repository of National Center for Biotechnology Information (NCBI) [15]. Based on the platform GPL570 Affymetrix Human Genome, the dataset with accession number GSE49515 [16] was retrieved. It contains a total of 26 samples of which 3 belongs to gastric cancer, 10 to HCC, 10 from normal patient, and last 3 from pancreatic cancer, from which we chose 20 samples including 10 HCC and 10 normal samples for our analysis; the gene expression data used are of

peripheral blood mononuclear cell (PBMC) from HCC patients to construct gene regulatory network, and PBMC is a noninvasive method to determine gene signature for detection of early human HCC [16].

2.2 DEGs Screening

The collected data were then organized into a matrix using MS Excel, where each row represents a gene and each column contains sample type. A function `matstest` performs a two-sample t-test to evaluate differentially expressed genes, and `mafdr` to estimate FDR of DEGs was used in MATLAB to infer the GRN [17]. The threshold of t-test and FDR is selected for p -value 0.01.

2.3 Network Construction and Analysis

DEGs are taken as node of network, and edges of the network are defined if the correlation is more than threshold values between any two genes for their gene expression data samples. A function called `corrcoef`, based on the Pearson correlation coefficient, was used for determining the relation among the DEGs in MATLAB [18]. Correlation coefficient is defined as

$$r = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y}$$

where $\text{cov}(x, y) = \sum (x - \bar{x})(y - \bar{y})$, $\sigma_x = \sqrt{\sum (x - \bar{x})^2}$ and $\sigma_y = \sqrt{\sum (y - \bar{y})^2}$ such that \bar{x} , \bar{y} are means of x , y , respectively.

Network analyzer is a plug-in in Cytoscape for performing topological analysis of directed as well as undirected network [19]. Network analyzer calculates and displays a broad set of topological parameters that include the number of nodes, edges, network diameter, connected components, radius, density, heterogeneity, and centralization.

2.4 Identification of Modules

One of the goals of this study is to detect potential modules in the co-expressed gene regulatory network for overall behavior associated with HCC and matched adjacent normal sample. To identify modules in the network and functional annotation of gene list, we used online tool GraphWeb [20], based on Markov's algorithms with default settings.

2.5 *Enrichment of Modules*

The modules obtained using GraphWeb tool was then enriched using online application WebGestalt [21, 22]. The DEGs in modular network were then translated into biological insight. Identifying the pathways where genes of modules are involved is one of the most central tasks of this study.

WebGestalt was used to organize genes (of module) in a table based on the KEGG pathways. These tables can predict the KEGG pathways, associated with modular genes' set and number of genes involved in each pathway. Additionally, KEGG table also displays the p -values which can be used for the inference of enrichment of each pathway.

3 Results and Discussion

3.1 *Differential Expressed Genes Screening*

The complementary DNA microarray analysis of the HCC and normal samples included 54,675 genes. A total of 1932 genes were selected as the DEGs between both the samples using t -test by taking threshold 0.01 p -value.

3.2 *Gene Regulatory Network of DEGs and Topological Analysis*

We have constructed network by taking node as differential expressed genes (DEGs) which have been screened by specific method and considering interactions (edges) among DEGs defined according to correlation coefficient between DEGs. Therefore, after taking minimum 0.95 cutoff criterion of existing edge between two DEGs, we got a network of 1932 nodes with 21,057 edges shown in Fig. 1. The network has no self-loop, multi-edges, and sparse network because the density of graph is too low and is equal to 0.011. The other several topological characteristics of constructed network can be seen in Table 1.

In the network, several DEGs' expressions are correlated with n number of other DEGs in the network; this n is hypothesized degree of node (DEG) in the network. Therefore, Table 2 is showing the highest degree node (genes) and their local characteristics. These genes are called hubs of the networks [23], and their important role is in the biological functions. Thus, the functions of those hubs are as follows:

ETNK1—the product of highest degree gene is an ethanolamine kinase that operates in the first step of phosphatidylethanolamine synthesis pathway and can be a rate controlling. This is very specific for ethanolamine phosphorylation.



Fig. 1 Gene regulatory network of differentially expressed genes between normal and HCC having 1932 nodes 21,057 edges

Table 1 Global characteristics of gene regulatory network of DEGs

| | |
|----------------------------|--------|
| Clustering coefficient | 0.334 |
| Connected component | 1 |
| No. of nodes | 1932 |
| Network diameter | 13 |
| Network radius | 7 |
| Network centralization | 0.125 |
| Characteristic path length | 3.956 |
| Avg. no. of neighbors | 21.798 |
| Network density | 0.011 |
| Network heterogeneity | 1.584 |
| Self loops | 0 |
| Multi edge node pairs | 0 |

Table 2 Top ten genes with high degrees

| Gene | Degree of nodes | Gene name | Clustering coefficient | Average shortest path length |
|---------|-----------------|--|------------------------|------------------------------|
| ETNK1 | 265 | Exosome component 3 | 0.18285781 | 2.48380463 |
| ST8SIA4 | 240 | Ethanolamine kinase 1 | 0.26614365 | 2.63598972 |
| SON | 233 | ST8 alpha-N-acetyl-neuraminide alpha-2,8-sialyltransferase 4 | 0.23719958 | 2.60257069 |
| UTP14C | 227 | General transcription factor IIH, polypeptide 1.62 kDa UTP14, IB small nucleolar | 0.28876067 | 2.67763496 |
| METTL13 | 218 | Ribonucleoprotein, homolog C (yeast) | 0.29104796 | 2.6781491 |
| GTF2H1 | 214 | GTPase, IMAP family member 6 | 0.2945345 | 2.63496144 |
| AP4E1 | 213 | Tripartite motif containing 27 | 0.32604976 | 2.67506427 |
| GIMAP6 | 201 | Methyltransferase like 13 | 0.31258706 | 2.72699229 |
| EXOSC3 | 199 | Adaptor-related protein complex 4, epsilon 1 subunit | 0.27980311 | 2.74858612 |
| TRIM27 | 195 | SON DNA binding protein | 0.26500432 | 2.65604113 |

ST8SIA4—encodes a protein which catalyzes the poly-condensation of α -2, 8-linked sialic acid which is required during the synthesis of polysialic acid (PSA). PSA works as modulator of neural cell adhesion molecule (NCAM1) particularly in adhesive properties.

SON—the outcome of this gene is protein which contains multiple repeats; the protein binds RNA and helps in pre-mRNA splicing specifically of transcripts with poor locations. The protein also binds DNA which belongs to hepatitis B virus and represses its core promoter activity.

UTP14C is a protein-coding gene, involved in ribosome biogenesis pathway in eukaryotes.

METTL13 involves in methyltransferase activity.

GTF2H1 is a protein coding gene according to GO annotation this gene involves in protein kinase activity, RNA polymerase II carboxy terminal domain kinase activity. The gene activity is also related to infectious disease and apoptotic pathways in synovial fibroblasts.

AP4E1 this results to the large protein family. These are the components of heterotetrametric adaptor protein com-complexes that play crucial roles in the pathways of secretory and endocytic via mediating vesicle formation and integral protein sorting.

GIMAP6 encodes a member of the GTPases for immunity-associated proteins (GIMAP) family. The resultant proteins contain GTP binding having coiled-coil

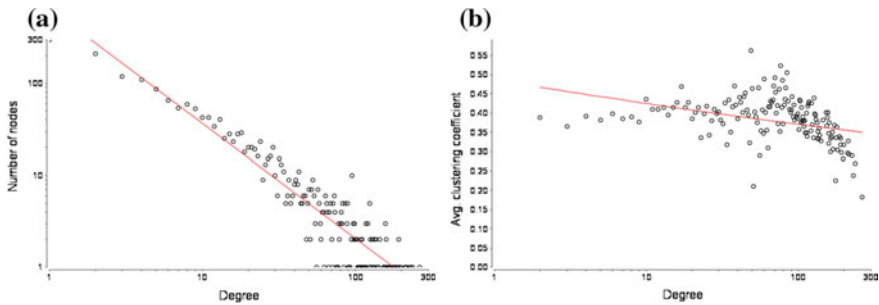


Fig. 2 **a** Node–degree distribution of DEGs and **b** average clustering coefficient distribution of DEGs

motifs which may play roles during the regulation of cell survival. Deregulation of this gene may involve in nonsmall cell lung cancer.

EXOSC3 refers to exosome component 3 which is a protein-coding gene. The genes involved in the pathway are transport into Golgi, subsequent modification, and unfolded protein response. The diseases related to this gene are pontocerebellar hypoplasia 1 and pontocerebellar hypoplasia 1b.

TRIM27 means tripartite motif-containing 27 is a gene that codes a protein which involves in thyroid carcinoma-related disease. GO identification related to this gene reveals its activity as nucleic acid binding and ubiquitin protein transferase activity. A very common global characteristic of network is degree distribution of network. It has found that most of the realistic network follows scale-free properties [13]. Figure 2a shows log-log plot degree versus number of nodes. The node–degree distribution plot is also showing a scale-free behavior like other network. Fitting in a power law $P(k) = ak^{-r}$, we found $r = 0.949$. Overall, the plot indicates the heterogeneity of degree in network, implying the co-expression of number of genes is heterogeneous, i.e., few gene expressions are correlated to large gene, and large a number of genes are just correlated to each other or few.

Clustering coefficient (CC) indicates clustering of network. It is defined for a node and gives ratio of number of edge among neighbors of i th node and the total possible edge with neighbors of i th nodes. The coefficient value 1 depicts the high-clustered neighbors of a node and 0 depicting no cluster for i th node. Average clustering coefficient gives the global characteristic of clustering. Figure 2b depicts the average clustering distribution of CC of k -neighbor node. Here, we have calculated average clustering coefficient for each degree, which gives the hierarchal structure of network [14], and inhomogeneity of clustering coefficient among the different degree nodes in the network. The GRN of DEGs every cluster can be seen in figure which follows the $c(k) k^{-a}$ where $a = 0.22$. Therefore, the network is not showing such hierarchical behavior as defined by Ravasz et al. [14]. But it is sure that clustering coefficient is inhomogeneity distributed and it is not well organized of modules but somehow organizations of modules are there.

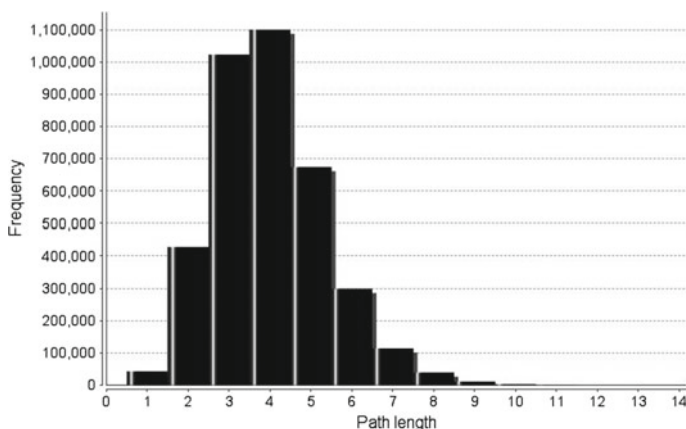


Fig. 3 Shortest path length distribution

According to the literature of network analysis, most of the empirical network follows the small world behavior [24] and GRN with average shortest path length 3.956. The distribution of path length in greatest integer form versus number of pairs of that path can be seen in Fig. 3.

Functional Analysis of Significant Modules

To investigate the modular composition within the gene regulatory network, we employed Markov algorithm and select top 20 modules (Table 3) based on node size and module density. Among them, we further chose top five modules for our study.

Module 1: The first module has 461 genes with 11,840 connections with themselves as shown in Fig. 4a. The predicted function of DEGs of module 1 distribution is based on gene ontology (biological process and molecular function) and KEGG pathway. The output depicts that major part of DEGs are involved in metabolic process. Past research about cancer has shown that it is a heterogenetic disorder [25], and also the metabolites formed as a result of central dogma are involved in cellular process. In past year's metabolomic-based study has revealed that a major fraction of metabolic function is deregulated in case of cancer formation [26]. The top ten genes with high degree are also present in first module. Genes such as ETNK1, ST8SIA4, SON, UTP14C, METTL13, GTF2H1, AP4E1, GIMAP6, EXOSC3, and TRIM27 show the high activity within the modular network. About the molecular function, 323 genes have binding function, 236 genes specifically have protein binding function, indicating the selective interaction of a molecule with one or more specific sites on other molecule. The KEGG pathway enrichment associated with first module shows the participation of DEGs in RNA transportation, metabolism, spliceosome, and toll-like receptor signaling pathway due to their contribution to many biological processes such as immune responses, healing of wounds, and carcinogenesis [27].

Module 2: The second module has 100 genes and 694 interactions shown in Fig. 4b. Biological process related to second module has shown that major fraction

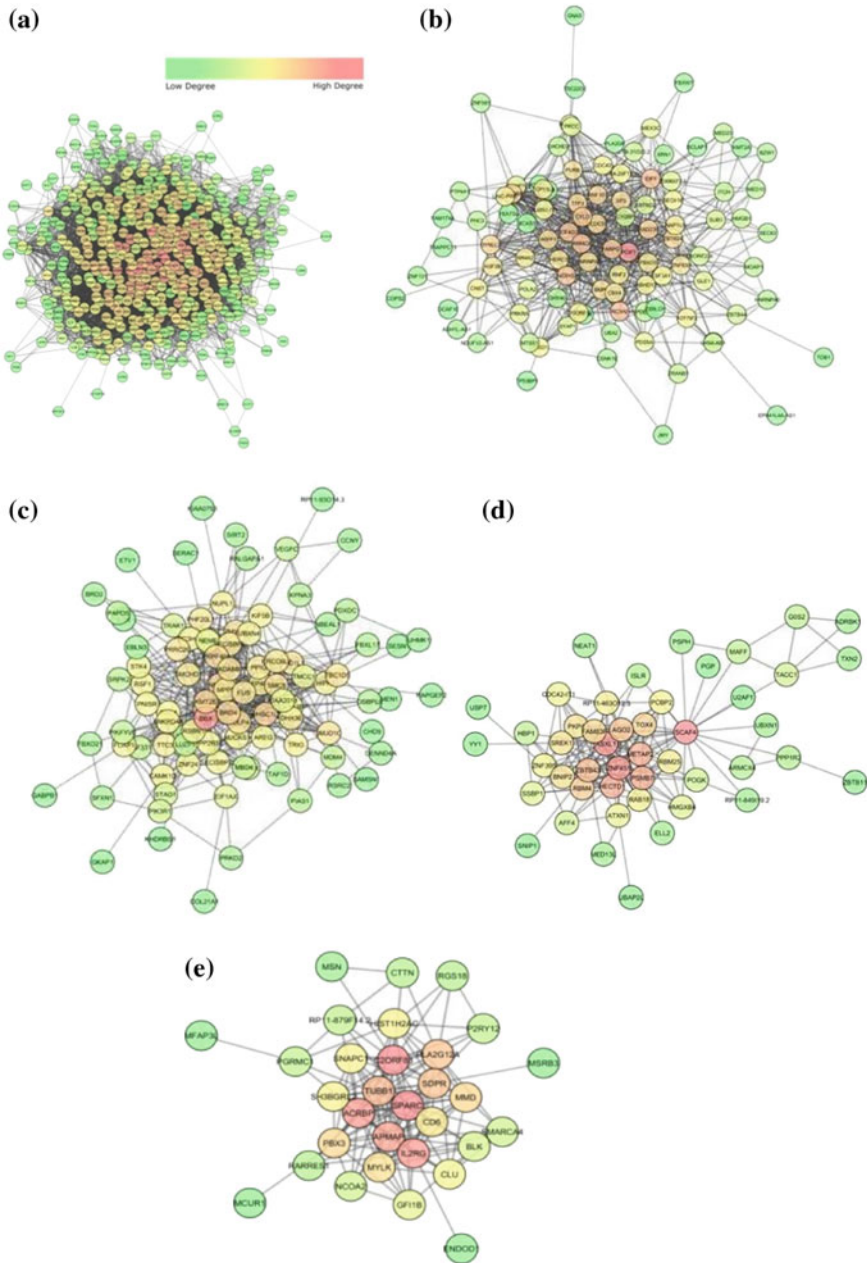


Fig. 4 Interactive networks of a module 1, b module 2, c module 3, d module 4, and e module 5

Table 3 Top 20 modules found in DEGs

| Module | #Nodes | #Edges | Density (%) |
|--------|--------|--------|-------------|
| 1 | 461 | 11,840 | 11.20 |
| 2 | 99 | 684 | 14.10 |
| 3 | 95 | 485 | 10.90 |
| 4 | 47 | 184 | 17.00 |
| 5 | 31 | 146 | 31.40 |
| 6 | 18 | 30 | 19.60 |
| 7 | 15 | 25 | 23.80 |
| 8 | 14 | 22 | 24.20 |
| 9 | 13 | 18 | 23.10 |
| 10 | 13 | 35 | 44.90 |
| 11 | 13 | 28 | 35.90 |
| 12 | 12 | 18 | 27.30 |
| 13 | 11 | 16 | 29.10 |
| 14 | 11 | 12 | 21.80 |
| 15 | 11 | 18 | 32.70 |
| 16 | 10 | 17 | 37.80 |
| 17 | 10 | 10 | 22.20 |
| 18 | 9 | 13 | 36.10 |
| 19 | 9 | 12 | 33.30 |
| 20 | 8 | 13 | 46.40 |

is involved in regulatory processes such as regulation of cellular biosynthetic process, macromolecule biosynthetic process, macromolecule metabolic process, gene expression, regulation of transcription, and gene expression. Apart from regulatory process, some genes are participating in protein modification by addition or removal of small protein. Most of the DEGs are associated in protein binding, DNA binding, and nucleic acid binding molecular function. The DEGs of second module in KEGG pathway depicts that, though less 4 genes Ubiquitin mediated proteolysis, 3 genes in GnRH signaling pathway.

Module 3: Third module has 95 genes with 484 interactions. Figure 4c has shown that 11 genes are participating in cell cycle process, 8 in chromatin modification, and 5 in cell cycle arrest. The KEGG pathway output of third module showed that its genes are involved in cancerous pathway, hepatitis C, glioma.

Module 4: The fourth module has 47 genes which are highly correlated to each other having 183 edges as shown in Fig. 4d. This module resembles the first module; the significant biological processes are broadly involved in metabolic processes such as macromolecular metabolic process, gene expression, and regulation of transcription. These are some processes which are differentially expressed either up- or down-regulated. In KEGG pathway 2 genes each in spliceosome and metabolic

pathway. ZNF451 is the most active gene within the modular network which may be involved in transcriptional regulation.

Module 5: The fifth module has 31 nodes and 145 edges as shown in Fig. 4e. There was no significant biological process found in this module. SPARC is the most active gene which encodes a protein required to collagen in bone for calcification, also involved in formation of extracellular matrix. Additionally, a study shown by Segat et al. that SPARC gene polymorphism is associated with HCC susceptibility in Italian patients. Product is correlated to tumor suppression [28].

4 Conclusion

This study used a graphical approach for analysis of gene expression data of normal and HCC samples. The co-expression network of DEGs follows the scale-free as well as hierarchical organization, as followed by other biological networks. Thus, the interactions among DEGs of resulted network follow heterogeneity due to specific nature of DEGs, whereas in the random network it follows homogeneity. This network also following the modular hierarchical nature which represents low-degree DEGs is much modular than higher-degree DEGs due to cohesively group of genes which are functioning for conclusive response of particular function.

The second interesting point is the modules of DEGs, because it is assumed that any individual gene does not play major role in functioning of cancer but the group of genes in coordinated fashion to bring about cellular function. Therefore, we have detected modules that are associated with specific purpose in cancer. The first and fourth modules involved mainly in metabolic process indicating the unhealthy state of cells, possibly cancer. Recent year's metabolomic-based study has revealed that a major fraction of metabolic function is deregulated in case of cancer [26]. The top ten genes are with respect to their degrees (from higher to lower) within the network, such as ETNK1, ST8SIA4, SON, and UTP14C.

METTL13, GTF2H1, AP4E1, GIMAP6, EXOSC3, and TRIM27 are seemingly playing active role in gene regulatory network which are also part of module 1. The second module is taking part into regulatory processes. The third module has 11 genes involve in cell cycle process, 8 in chromatin modification, and 5 in cell cycle arrest. Within fifth module, SPARC is the most active gene which regulates growth of cell by Segat et al. that SPARC gene polymorphism is associated with HCC susceptibility [28].

These findings may be beneficial for those researchers who are working in the field of biomarker and diagnosis of cancer directly from PBMC samples. However, many studies have done, but group of gene that is co-expressed simultaneously is given by this study.

Acknowledgements This work is supported by the project grant received from SERB (File No.-EEQ/2016/000509), DST, Govt. of India, and was carried out at the Centre for Interdisciplinary Research in Basic Sciences, Jamia Millia Islamia, New Delhi, India.

References

- Jacques, F., Soerjomataram, I., Dikshit, R., Eser, S., Mathers, C., Rebelo, M., et al. (2015). Cancer incidence and mortality worldwide: Sources, methods and major patterns in globocan 2012. *International Journal of Cancer*, 136(5).
- Kew, M. C. (2010). Epidemiology of chronic hepatitis b virus infection, hepatocellular carcinoma, and hepatitis b virus-induced hepatocellular carcinoma. *Pathologie Biologie*, 58(4), 273–277.
- Mittal, S., & El-Serag, H. B. (2013). Epidemiology of HCC: Consider the population. *Journal of Clinical Gastroenterology*, 47, S2.
- El-Serag, H. B. (2012). Epidemiology of viral hepatitis and hepatocellular carcinoma. *Gastroenterology*, 142(6), 1264–1273.
- El-Serag, H. B. (2004). Hepatocellular carcinoma: Recent trends in the United States. *Gastroenterology*, 127(5), S27–S34.
- Bruix, J., & Sherman, M. (2005). Management of hepatocellular carcinoma. *Hepatology*, 42(5), 1208–1236.
- He, T. L., Zheng, K. L., Li, G., Song, B., & Zhang, Y. J. (2014). Identification of typical miRNAs and target genes in hepatocellular carcinoma by DNA microarray technique. *European Review for Medical and Pharmacological Sciences*, 18(1), 108–116.
- Plentz, R. R., Boozari, B., & Malek, N. P. (2014). Guideline compliant diagnostics of hepatocellular carcinoma. *Der Radiologe*, 54(7), 660–663.
- Hartwell, L. H., Hopfield, J. J., Leibler, S., & Murray, A. W. (1999). From molecular to modular cell biology. *Nature*, 402, C47–C52, 26(27), 59–63.
- Mitra, K., Carvunis, A.-R., Ramesh, S. K., & Ideker, T. (2013). Integrative approaches for finding modular structure in biological networks. *Nature Reviews Genetics*, 14(10), 719–732.
- Lau, W. Y., Lai, P. B. S., Leung, M. F., Leung, B. C. S., Wong, N., Chen, G., et al. (2001). Differential gene expression of hepatocellular carcinoma using cDNA microarray analysis. *Oncology Research Featuring Preclinical and Clinical Cancer Therapeutics*, 12(2), 59–69.
- Zhuang, L., Wu, Y., Han, J., Ling, X., Wang, L., Zhu, C., & Fu, Y. (2014). A network biology approach to discover the molecular biomarker associated with hepatocellular carcinoma. *BioMed Research International*, 2014.
- Barabási, A. L., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286(5439), 509–512.
- Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N., & Barabási, A.-L. (2002). Hierarchical organization of modularity in metabolic networks. *Science*, 297(5586), 1551–1555.
- Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., et al. (2012). Ncbi geo: Archive for functional genomics data setsupdate. *Nucleic Acids Research*, 41(D1), D991–D995.
- Shi, M., Chen, M.-S., Sekar, K., Tan, C.-K., Ooi, L. L., & Hui, K. M. (2014). A blood-based three-gene signature for the non-invasive detection of early human hepatocellular carcinoma. *European Journal of Cancer*, 50(5), 928–936.
- Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3), 479–498.
- Adler, J., & Parmryd, I. (2010). Quantifying colocalization by correlation: The Pearson correlation coefficient is superior to the Mander's overlap coefficient. *Cytometry Part A*, 77(8), 733–742.
- Assenov, Y., Ramírez, F., Schelhorn, S.-E., Lengauer, T., & Albrecht, M. (2007). Computing topological parameters of biological networks. *Bioinformatics*, 24(2), 282–284.
- Reimand, J., Arak, T., & Vilo, J. (2011). g: Profiler web server for functional interpretation of gene lists (2011 update). *Nucleic acids research*, 39(suppl 2), W307–W315.
- Zhang, B., Kirov, S., & Snoddy, J. (2005). Webgestalt: An integrated system for exploring gene sets in various biological contexts. *Nucleic Acids Research*, 33(suppl 2), W741–W748.
- Wang, J., Duncan, D., Shi, Z., & Zhang, B. (2013). Web-based gene set analysis toolkit (webgestalt): Update 2013. *Nucleic Acids Research*, 41(W1), W77–W83.

23. Han, J.-D. J., Bertin, N., Hao, T., Goldberg, D. S., Berriz, G. F., Zhang, L. V., et al. (2004). Evidence for dynamically organized modularity in the yeast protein–protein interaction network. *Nature*, 430(6995), 88–93.
24. Albert, R., & Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(1), 47.
25. Avila, M. A., Berasain, C., Sangro, B., & Prieto, J. (2006). New therapies for hepatocellular carcinoma. *Oncogene*, 25(27), 3866–3884.
26. Ramesh, V., & Ganesan, K. (2016). Integrative functional genomic delineation of the cascades of transcriptional changes involved in hepatocellular carcinoma progression. *International Journal of Cancer*, 139(7), 1586–1597.
27. Eiró, N., Altadill, A., Juárez, L. M., Rodríguez, M., González, L. O., Atienza, S., et al. (2014). Toll-like receptors 3, 4 and 9 in hepatocellular carcinoma: Relationship with clinicopathological characteristics and prognosis. *Hepatology Research*, 44(7), 769–778.
28. Segat, L., Milanese, M., Pirulli, D., Trevisiol, C., Lupo, F., Salizzoni, M., et al. (2009). Secreted protein acidic and rich in cysteine (sparc) gene polymorphism association with hepatocellular carcinoma in italian patients. *Journal of Gastroenterology and Hepatology*, 24(12), 1840–1846.

Comparative Analysis of Various Techniques of DDoS Attacks for Detection & Prevention and Their Impact in MANET



Neha Singh, Ankur Dumka and Rakesh Sharma

Abstract Wireless nodes all connected together logically make a mobile ad hoc network (MANET). Fundamentally, MANET has no infrastructure which means all the connections are formed without the help of or use of one or more administration which is centralized. A mobile ad hoc network (MANET) can be termed as a network which is spontaneous and which can be set up without any fixed infrastructure. The main goal of denial-of-service (DoS) attack is to make the server and other resources too busy so that request for any information or resource is either not served at all or there is an unexpected delay in response time. When a DoS attack is converted into its severe form, then it is known as distributed denial-of-service (DDoS) attack. A DDoS attack uses several machines to launch the attack by making services block for genuine users. A DDoS attack which can be in form of a bandwidth depletion and resource depletion prevents the legitimate users by flooding the victim's network with unwanted traffic in form of packets. Various techniques have been implemented to secure the channel, but still security is a major issue over the network. The aim of this research work is to present a comparative analysis of various proposed tools and techniques for detection and prevention of DDoS attacks.

Keywords Mobile ad hoc networks (MANET) · Denial of service (DoS) · Distributed denial-of-service (DDoS) · Dynamic nodes · Blackhole attacks

N. Singh (✉)

Uttarakhand Technical University, Uttarakhand, India
e-mail: singh.neha773@gmail.com

A. Dumka

Graphic Era, Deemed to be University, Dehradun, India
e-mail: ankurdumka2@gmail.com

R. Sharma

Government of Jharkhand in Higher & Technical Education, Jharkhand, India
e-mail: colonelrakesh_ajm@yahoo.co.in

1 Introduction

Mobile ad hoc network (MANET) [1] consists of various nodes which are mobile, and dynamically these nodes constitute some multihop communication which leads to denial-of-service (DoS) attack, selfish misbehaving, etc. A secure communication is a challenging task in MANET. Basically, old mechanisms of security which are designed for infrastructure networks are not capable for MANET as of its characteristics. Also, MANET suffers from a large number of security threats due to its nature of dynamically changing topology and also absence of centralized administration.

1.1 Flooding Attack in MANET

It [2, 3] is known to be the attack commonly found in mobile ad hoc network. The main objective of such attacks is to drain the resources available in the network like its bandwidth, node processing power, battery power, and to waste network resources by engaging it aimless and just degrade its performance. Many of the resources got wasted while trying to do route generation for the destination that not even exist.

The attack whose objective is to only flood the network is termed as route request flooding attack (RRFA). Large numbers of packets are generated during this attack in form of RREQ, and network will not become to transmit packets. In this attack, random IP address is selected according to the scope.

1.2 Effects of Flooding Attack

This flooding attack always affects the network node and also decreases the performance and efficiency of reactive protocols in the following manner:

- I. Buffer performance degraded: The routing protocol is using buffer to store/buffer data packets when it will do route searching, and due to many packets the buffer may overflow. Also, genuine data packets which application layer is generating may become out of reach and authenticated packets can be substituted by unreachable packets [4, 5].
- II. Wireless interface performance degraded: Main cause is the buffer overflow which used by the card, i.e., wireless interface network card. As large number of data packets asking for RREQ cause the overflow to occur which results in dropping of genuine data packets.
- III. RREQ packets performance degraded: The entire network is broadcast with large number of RREQ packets; more MAC layer collision will occur due to large building of packets of RREQ in the network [6] which results into traffic in the whole network which also causes delay of packets.

IV. Degrade the performance of lifetime in MANET: Large number of useless RREQ transmission will decrease the network lifetime, and also it can increase overhead of authenticating large amount of RREQs.

DoS attack: Denial-of-service (DoS) attack [3] does flooding of a server with several data packets with the help of one computer. The only objective of this type of attack is to make server bandwidth overloaded. The severe form of DoS is DDoS attack which prevents legitimate user to avail services by using multiple machines. To attack Internet resources, it is an active and strong method.

The DoS problem can be added as many-to-one dimension, and that is the reason that prevention or detection schemes or methods are complicated. To initiate an attack, first an attacker connect within a network which consist of several nodes. Under the control of an attacker, zombies carry out attacks on the nodes. These zombies are basically of two types, i.e., masters and slaves. Now the attacker node signals masters to attack, and then masters motivate their slaves to attack.

DDoS attacking patterns and a defense scheme: Overview of IEEE 802.11: To minimize the collision of packets in a network, Distributed Coordination Function (DCF) of IEEE 802.11 describes how to minimize them by using carrier-sense multiple access with collision avoidance (CSMA/CA) [7]. Randomly, the node selects any back-off value i which will transmit a packet from the given set $\{0, 1 \dots CW\}$ in which CW represents the size of the contention window that will wait for i idle time slots which transfer the packet.

For engaging the channel of network, the nodes send Request to Send (RTS) and Clear to Send (CTS) packets before transmission [7]. For completion of data transmission, packets like RTS and CTS indicate how much time they require for the channel utilization. To determine how much time a node would take for transmission over the channel, both RTS and the CTS are required to adjust their Network Allocation Vector (NAV).

Sometimes due to lack of CTS for RTS and for ACK for sent data, the transmission is termed as unsuccessful and which causes value of CW as doubled and data packet is transmitted again. For RTS, the maximum value of retransmission is always 7 and for data it is set to 4. On the contrary, if the transmission is fully successful, then the host again sets the value of CW to CW_{min} [8]. Idle time spaces, distributed inter-frame space (DIFS) and short inter-frame space (SIFS), are added between all the frames.

Attacking principles: Channel is said to be the main resource for wireless networks [1, 9], in which only one node at a time has the access from all the nodes the wireless network shares whereas in the wired network congestion always occur due to lot of data traffic at various bottleneck link, also in MANET congestion occurs due to several nodes in form of aggregation. Most of the times, the attacking nodes always defeat the victim node and occupy the channel.

One more reason of such type of attacks is the methods or routing protocols used by nodes of MANET which do not provide good traffic controlling mechanism in form of traffic filters, allocation of traffic, and QoS. As the studies suggest nodes or radio distance between nodes must be more to make the concurrent data transmitted.

The aggregate number [10] of information parcels that can be at the same time transmitted for one-jump increments directly with the aggregate region of the ad hoc network. In the event that the hub thickness is steady, at that point the aggregate one-jump limit is $O(n)$, and n is the aggregate number of hubs in the system.

The gaps between the source and the goal will increase accordingly as with time system starts developing. With the spatial distance across of the system [8, 11], the normal way length develops, which calculates the square foundation of the range, i.e., $O(pn)$. With this study, we can say that the conclusion till the end is that throughput accessible to every hub is $O(1/pn)$ and the end-to-end limit is generally $O(n/pn)$.

2 Literature Survey

Huge works have been done in securing the promotion arranges. Some inquires about characterized technique for secure directing however secure steering likewise cannot ready to deal with the flooding assault. The information flooding assault causes denial-of-service (DoS) assaults by flooding numerous information parcels.

Be that as it may, there is a couple of existing barrier frameworks against information flooding assaults. Besides, the current plans may not ensure the Quality of Service (QoS) of burst activity since interactive media information is typically blasted.

This paper shows surge assault which help in disabling IP broadcast; it is a procedure which is recommended to keep a DDoS assault. This proposed method has the ability to shutdown the distributed denial-of-service attacks [12]. It can be observed that the proposed arrangement gives a better result than the existing plans.

The significant test of mobile ad hoc network is basically its security, since this does not come with a brought together control. Mobile ad hoc network likewise consists of a sensor arrangement; therefore, it additionally confronts the issue confronted through sensor systems. Various number of security assaults exist in MANET, and DDoS (distributed refusal of administration) [9] is the most challenging one among them. Our principle point is to see the impact of DDoS.

Diverse components have been proposed for utilizing different cryptographic systems to countermeasure these assaults against MANET. Such components are not reasonable for MANET [1] asset imperatives, i.e., constrained transfer speed and battery control since they acquaint overwhelming activity stack with trade and confirming keys. In this examination, the present issues of security against MANET are researched. Especially, the analysts have analyzed distinctive DDoS assaults and some other recognition strategies like profile-based discovery particularly based on identification and in addition existing answers to ensure MANET conventions.

In this paper, disseminated refusal of administration assaults (DDoS) is displayed which are assaulted on versatile specially appointed system and exhorted way to deal with distinguish DDoS assault and give legitimate answers to amplify arrange execution and assets through correlation [13] of various system parameters.

The introduced approach depends on the examination of across the board data transfer capacity assaults, with concentrate on distributed denial-of-service (DDoS) assaults, which are to a great degree hazardous, difficult to recognize and testing to avert. DDoS speaks to a planned movement of a gathering of assailants intending to counteract authentic clients the entrance to arrange assets. Interruption avoidance frameworks (IPS) are predominantly considered as expansions of interruption discovery frameworks (IDS) with a reason to effectively counteract and square interruptions that are identified by IDS [14]. The introduced IPS show that it depends on the examination of the legal investigation report created by IDS consolidated into the system security observing framework.

The primary goal of this is similar investigation of different sorts of DDoS assaults and different discovery strategies and also resistance components like disable IP communicate recognition system, profile-based identification, and counteractive action of DoS assault utilizing target client conduct and existing answers for ensure MANET conventions. With the assistance of these methods, number of [2] crashes and bundle conveyance proportion are additionally assessed by contrasting two counteractive action procedures existing anticipation strategy and proposed avoidance system. The proposed system expands PDR (Packet delivery ratio) and decreases number of crashes.

Huge endeavors have been made toward influencing ad hoc to arrange some more secure and DDoS free. In this study, we examine how different discovery parameters all together function as a solitary and productive strategy to identify different DDoS assaults for MANET. Later in this study [3], a strategy to anticipate DDoS assaults in MANET is likewise introduced which help in keeping the assaults to convey in the system and did not permit them in the system.

In this paper, we have a tendency to look at the conduct of different assaults result in organize. In this overview, we significantly feature the conduct of various assaults with particular consistence of Jamming assault and guard conspires in MANET. The multipath steering plans are likewise talked about to enhance the system execution in arrange however condition is that sticking condition is conceivable to happen by assailant [7]. In nearness of assailant security plot are dependably gives the safe way then in multipath steering the likelihood of secure directing is improved in nearness of aggressor and security conspire.

Military battlefields and various commercial applications like traffic systems have been using MANET which come across as an emerging technology and has great strength [15]. With no centralized controller, it acts as an infrastructure less. As no central controller exists, therefore security is considered as a major challenge in wireless mobile ad hoc network. DDoS can be termed as one of the major attacks from the various attacks in MANET.

Detection and prevention of mobile ad hoc network considered to be the prime objective of the study and keep it away from unwanted attacks which cause power draining of the resources. In this paper, vampire attack has been deployed using flooding through nodes with high battery capacity. Also to detect affected nodes various energy consumption and capacity observation methods has been suggested.

Additionally [10], a prevention technique shutdown the attacker node forcefully and help in establishing the communication.

MANET becoming day by day more prone to DDoS attacks, draining of resources, blackhole, grayhole, etc. Due to this reason, security issues need to be look on as soon as possible and allow only authenticate and authorized users to access the data provided on the network. In this study [11], a technique is proposed which help in preventing from the attacks. The effective technique supported by showing results using simulation in GloMoSim which also combined with Parsec compiler on a windows platform.

Privacy and security have turned into an irreplaceable matter of consideration in the VANET networks, which is helpless against numerous security dangers nowadays. One of them is the denial-of-service (DoS) assaults, where a vindictive hub fashions a substantial number of phony characters, i.e., Internet Protocol (IP) delivers so as to disturb the correct functioning of meaningful exchange of information between any two quickly moving vehicles. In this study, different levels of denial-of-service assault in VANET [4] are talked about and different methodologies that relieve the effect of DoS, Jamming and distributed DoS assaults are overviewed, and a basic plan is proposed to defeat DoS.

The proposed approach depends on the examination and examinations of transfer speed assaults that predominantly concentrate on DDoS that is genuinely a savage test and is hard to recognize, and diminishes the execution of the system. DDoS incorporates a gathering of assailant hubs and focuses on the casualty to keep the authentic clients from getting to the system administrations and assets.

Interruption avoidance frameworks are the techniques that are dealt with as additions of the interruption recognition framework to effectively guard and keep the interruptions that are distinguished by the identification strategies of the IDS [5]. The report that is produced by the IDS in the wake of examining the report of the criminological examination is the base of the proposed methodology.

Distributed refusals of administration (DDoS) assaults are characterized as assaults that are propelled by an arrangement of malevolent substances toward a hub or set of hubs. In this work, we propose an answer for keeping WSN from DDoS assault utilizing dynamic source steering (DSR) [6]. Vitality of concerned hubs has been utilized for recognition and anticipation of assault. QualNet 5.2 test system is utilized for usage of the proposed arrangement.

The correspondence in MANET works appropriately just if taking an interest hubs participate in steering with no malevolent expectation. Since a MANET does not have any foundation, sudden flooding would bring about execution corruption and would bring about the end of the correspondence occurring [16]. This exploration paper investigates the effect of flooding on MANET.

In this paper, a novel arrangement was proposed to shield the OLSR convention starting hub confinement assault by utilizing the identical methodology worn by the pounce upon itself. Completely through broad testing, we make clear that (1) the proposed security averts supplementary than 95% of assaults, and (2) the overhead required extensively diminishes as the system measure increments in suspicion of it

is non-detectable [17]. To wrap things up, these activities recommend that this sort of determination can be stretched out to other comparable DoS assaults on OLSR.

In this paper, a hub trust count procedure is proposed which computes the trust estimation of every hub and applies fluffy rationale to identify wormhole, black-gap (routing assault), and appropriated refusal [18] of administration assault (DDoS/flooding) in powerful condition.

This paper is a study on the issue of forswearing DDoS benefit (DoS) assaults and proposed approaches to recognize it. This elaboration is to investigate the powerlessness of refusal DDoS benefit assault. We found that conveying a productive interruption identification framework can fundamentally enhance its security and it can distinguish forswearing of administration assault before it influences the casualty.

In any case, because of the extraordinary qualities the current conventional interruption recognition framework cannot work appropriately in that condition. The concentration of this paper [19] is to examine recognition framework which can distinguish DoS assault with high assault location and low false alert rate to find the attackers.

A decent measure of research has been finished amid the current past for location and aversion of this sort of assault in order to keep up the execution and unwavering quality of the remote sensor systems [20]. In this exploration article, the effect of blackhole movement is assessed utilizing system fundamental parameters and a novel procedure is intended to distinguish and keep the blackhole assault in remote sensor systems.

The multijump nature of the message transmission carries alongside the genuine security concerns. The gadgets need to work longer for the mind boggling applications and different employments [21]. The assaults centered at depleting or devouring of the batteries, for example, flooding assault winds up plainly imperative to consider in the perspective of vitality preservation and improvement of the system life expectancy. This paper demonstrates the different examinations with respect to the DDoS flooding assault in the systems.

3 Comparative Study

A comparative study of the proposed tools and techniques used for distributed denial of service (DDoS) is shown in the following Table 1.

4 Conclusion and Future Work

Refusal of administration assault utilizes a conveyed structure which could be utilized to recognize and keep the authentic utilization of administration. Distinctive strategies of DoS attack in MANET have been surveyed. Point of this assault is

Table 1 Comparison of proposed tools and techniques for DDoS

| S.No | Title | Author | Published at | Tools | Technique |
|------|--|---|--|-----------------------------------|---|
| 1 | Detection and prevention of DDoS attack in MANET's using disable Ip broadcast technique [12] | Mukesh Kumar & Naresh Kumar | International Journal of Application or Innovation in Engineering & Management | GloMoSim | Detection and prevention technique |
| 2 | A relative study for detection and prevention of DDoS attacks [9] | Ms. Anjusree, S & Mrs. V. Praveena | International Journal of Innovative Research in Computer & Communication Engineering | NA | Detection method using tasteful and stateless |
| 3 | Attack prevention methods for DDoS attacks in MANETs [1] | Neeraj Sharma, B. L. Raina, Prabha Rani, Yogesh Chaba & Yudhvir Singh | Asian Journal Of Computer Science And Information Technology | NA | Prevention technique |
| 4 | Detection and prevention mechanism for TTL field tampering form of DDoS attack in MANET's [13] | Deepak Vishwakarma, Nitin Rathore & Anil Khandekar | International Journal of Computer Applications | NS-2.35 | Detection and prevention scheme |
| 5 | An approach for DDoS attack prevention in mobile ad hoc networks [14] | V. V. Timeenko | ELEKTRONIKA IR ELEKTROTECHNIKA | Ns-2(ver 2.34) in Linux fedora 16 | Prevention technique |
| 6 | DOS attack mitigation In MANET [2] | Er. Inakshi Garg, Er. Meenakshi Sharma | IJEDR | GloMoSim Tool | Mitigation technique |
| 7 | Explicit query based detection and prevention techniques for DDoS in MANET [3] | Neha Singh, Sumit Chaudhary, Kapil Kumar Verma & A. K. Vatsa | International Journal of Computer Applications | NA | Detection and prevention scheme |

(continued)

Table 1 (continued)

| S.No | Title | Author | Published at | Tools | Technique |
|------|--|--|--|------------------------|---------------------------------|
| 8 | Detection and prevention scheme against jamming attack in MANET [7] | Namrata Soni & Vinay Singh | International Journal of Computer Applications | NA | Detection and prevention scheme |
| 9 | A survey on intrusion detection system for DDoS attack in MANET [15] | Nshunguye Justin & Nitin. R. Gava | IJARCCCE | NA | Survey on DDoS attacks |
| 10 | Detection and prevention of vampire attack in MANET [10] | Anamika Garg & Mayank K Sharma | International Journal on Recent and Innovation Trends in Computing and Communication | NS-2, S/W version 2.35 | Detection and prevention scheme |
| 11 | An efficient scheme to prevent DDoS flooding attacks in mobile ad hoc network (MANET) [11] | Meghna Chhabra & B. B. Gupta | Research Journal of Applied Sciences, Engineering and Technology | GloMoSim | Prevention scheme |
| 12 | Approaches to reduce the impact of DOS and DDoS attacks in VANET [4] | Aaditya Jain & Divya Sharma | International Journal of Computer Science | NA | Survey |
| 13 | Preventive mechanism against DDoS attacks in MANET [5] | Tariq Ahamad & Abdullah Aljumah | International Journal of Advanced and Applied Sciences | NS-2 network simulator | Prevention algorithm |
| 14 | DDoS attack aware DSR routing protocol in WSN [6] | Raksha Upadhyaya, Uma Rathore Bhatta & Harendra Tripathi | Procedia Computer Science, Elsevier | QualNet 5.2 simulator | Defensive scheme |

(continued)

Table 1 (continued)

| S.No | Title | Author | Published at | Tools | Technique |
|------|---|---|---|------------|------------------------------------|
| 15 | Flooding attack on MANET—A survey [16] | C. M. Nalayini, Dr. Jeevaa Katiravan & Arvind Prasad. V | IJTRD | NA | Survey |
| 16 | Detection and prevention of DDoS attack using modern cracking algorithm [17] | S. G. Suganya & D. Prasanna | International Research Journal In Advanced Engineering And Technology | NA | Detection and prevention scheme |
| 17 | Detection of wormhole, blackhole and DDoS attack in MANET using trust estimation under Fuzzy logic methodology [18] | Ashish Kumar Khare, Dr. J. L. Rana & Dr. R. C. Jain | I. J. Computer Network and Information Security | NS-2, 2.34 | Detection scheme |
| 18 | A study and analysis of DoS attacks and prevention scheme [19] | Jagpal & Gaurav Garg | International Journal of Innovative Research in Computer and Communication Engineering | NA | Survey |
| 19 | Futuristic method to detect and prevent blackhole attack in wireless sensor networks [20] | Abdullah Aljumah & Tariq Ahamed Ahanger | International Journal of Computer Science and Network Security | NS 2 | Detection and prevention mechanism |
| 20 | Review on the flooding attacks in mobile ad hoc networks [21] | Jasmeen Mangat & Er. Jaspreet Kaur | International Journal of Advanced Research in Computer Science and Software Engineering | NA | Review paper |

endeavored to accomplish a reasonable perspective of the DDoS assault issue and discover more powerful answer for the issue.

DDoS attacks can make an arranged framework or administration accessible to honest to goodness clients and utilizations numerous machines over the network for keeping the assets and provide a guard plan to moderate the assault in remote Ad hoc systems.

By sending a large number of data packets in a huge volume to the target machine, a DDoS attack was launched with the cooperation of various numbers of host which are distributed all over the network. For detecting and preventing the DDoS attacks, constant research work is going on in MANET.

A review of various detection and prevention techniques with using various tools has been shown in this paper. In future, more researches can be done for improving MANET as the network is heterogeneous, mobile, scalable, and with the ongoing advancement in this field, it is fully composite.

References

1. Sharma, N., Raina, B. L, Rani, P., Chaba, Y., & Singh, Y. (2011). Attack prevention methods for DDoS attacks in MANETs. *AJCSIT*, 18–21.
2. Garg, E. I., & Sharma, E. M. (2016). DOS attack mitigation in MANET. *IJEDR*, 4(3), 769–773.
3. Singh, N., Chaudhary, S., Verma, K. K., & Vatsa, A. K. (2012, September). Explicit query based detection and prevention techniques for DDoS in MANET. *IJCA*, 53(2), 19–24.
4. Jain, Aaditya, & Sharma, Divya. (2016). Approaches to reduce the impact of DOS and DDoS attacks in VANET. *IJCS*, 4(4), 1–5.
5. Ahamad, Tariq, & Aljumah, Abdullah. (2017). Preventive mechanism against DDoS attacks in MANET. *International Journal of Advanced and Applied Sciences*, 4(5), 94–100.
6. Upadhyaya, R., Bhatta, U. R., & Tripathi, H. (2016). DDoS attack aware DSR routing protocol in WSN. *Procedia Computer Science*, 78, 68–74. (Elsevier).
7. Soni, N., & Singh, V. (2016, November). Detection and prevention scheme against jamming attack in MANET. *IJCA*, 154(11), 37–41.
8. Lovely. (2015, April). A review of DoS attack and defence scheme in MANET. *IJARCSSE*, 5(4), 1107–1112.
9. Anjusree, S., & Praveena, V. (2013, October). A relative study for detection and prevention of DDoS attacks. *IJIRCCCE*, 1(8), 1786–1792.
10. Garg, A., & Sharma, M. K. (2015, December). Detection and prevention of vampire attack in MANET. *IJRITCC*, 3(12), 6793–6798.
11. Chhabra, M., & Gupta, B. B. (2014, March 15). An efficient scheme to prevent DDoS flooding attacks in mobile. *Research Journal of Applied Sciences, Engineering and Technology*, 7(10), 2033–2039.
12. Kumar, Mukesh, & Kumar, Naresh. (2013). Detection and prevention of DDoS attack in MANET's using disable IP broadcast technique. *IJAEM*, 2(7), 29–36.
13. Vishwakarma, D., Rathore, N., & Khandekar, A. (2015, May). Detection and prevention mechanism for TTL field tampering form of DDoS attack in MANET's. *IJCA*, 117(15), 5–10.
14. Timcenko, V. V. (2014). An approach for DDoS attack prevention in mobile ad hoc networks. *Elektronika ir Elektrotechnika*, 20(6), 150–153.
15. Justin, N., & Gava, N. R. (2016, April). A survey on intrusion detection system for DDoS attack in MANET. *IJARCCCE*, 5(4), 1160–1163.
16. Nalayini, C. M., Katiravan, J., & Arvind Prasad, V. (2017). Flooding attack on MANET—A survey. *IJTRD*, ISSN: 2394–9333, 25–28.

17. Suganya, S. G., & Prasanna, D. (2017). Detection and prevention of DDoS attack using modern cracking algorithm. *IRJAET*, 3(2), 2004–2012.
18. Khare, A. K., Rana, J. L., & Jain, R. C. (2017). Detection of wormhole, blackhole and DDoS attack in MANET using trust estimation under Fuzzy logic methodology. *MECS*, 7, 29–35.
19. Jagpal, & Garg, G. (2017, May). A study and analysis of DoS attacks and prevention scheme. *IJIRCCE*, 5(5), 10379–10382.
20. Aljumah, A., & Ahanger, T. A. (2017, February). Futuristic method to detect and prevent blackhole attack in wireless sensor networks. *IJCSNS*, 17(2), 194–201.
21. Mangat, J., & Kaur, E. J. (2017, April). Review on the flooding attacks in mobile Ad Hoc networks. *IJARCSSE*, 7(4), 390–392.

A Comparative Study of Data Mining Tools and Techniques for Business Intelligence



G. S. Ramesh, T. V. Rajini Kanth and D. Vasumathi

Abstract Business intelligence (BI) is a collection of different frameworks and tools that convert the required raw data into meaningful information which may aid in supporting the decision-making process of the management. The present-day BI gives a reporting functionality to the identification of data groups, i.e., clusters useful for data mining techniques and business performance maintenance with predictive analysis in real-time BI applications. In fact, the core function of BI is to support the effective decision-making process. The BI frameworks are often known to business clients as decision support systems (DSSs) or reality-based supporting systems that they utilize to analyze the data and extract information from data sources. Through this research work, the authors aim to discuss various tools, approaches, and techniques for data mining that has support for BI. The research work also aims to describe the study as processes and procedures to systematically identify, counter, store, analyze, and explore data accessibility for making effective operations in business decisions. Different algorithms, methods, and techniques of BI are also highlighted along with varied types of applications with preferable implementation. The later part of the study discusses BI applications, which include operations of decision supporting frameworks, data management frameworks, query and reporting with online analytical processing (OLAP), forecasting apart from statistical analysis used in distributed BI applications. This research work uses visualized charts to explain the usage frequency of BI techniques with their performance comparison.

G. S. Ramesh (✉)
Department of CSE, VNR VJIET, Hyderabad, India
e-mail: ramesh_gadwal@yahoo.com

T. V. Rajini Kanth
Department of CSE, SNIST, Hyderabad, India
e-mail: rajinitv@gmail.com

D. Vasumathi
Department of CSE, JNTU-H, Hyderabad, India
e-mail: vasukumar_devara@yahoo.co.in

Keywords Business intelligence (BI) · Decision support systems · Online analytical processing (OLAP) · Data mining · Clustering · Reporting · Statistical analysis

1 Introduction

Business intelligence (BI) is a completely separate classification of various applications and technologies that combines and retrieves permissions before analyzing information in order to help business enterprise customers to make efficient business decisions. As suggested by its name, BI consists of the entire bulk of knowledge related to business implementation factors like customers, competitors, business partners, and economic environment with internal operations for effective business operations. BI provides readers the flexibility with some mathematical evaluations, data mining applications, and data analysis for making transactional data attribute presentation in business intelligence. In essence, business intelligence provides knowledge about data mining tools and methodologies that are critical for defining operations as well as allowing them to develop timely decisions.

2 Related Work

Several studies deal with the technological features considered necessary for signing the review pathway information. Gounder et al. [1] addressed the problem of workflow record control in exclusive businesses. In order to provide a simpler framework as an affordable solution for businesses, they recommend reducing the workflow record by saving the chosen and aggregated information. In order to achieve this objective, they increased the workflow specification by means of orthogonal condition mapping elements and corresponding actions for record monitoring. The model Mentor-lite utilizes record control as a workflow on the top of a light and portable kernel. A strategy for the monitoring record information in an allocated workflow control system has also been provided in Tom and Ghosh [2]. The contributors focused on the framework and querying of the historical past in a fully allocated framework that is in conformity with Object Management Architecture (OMA). As a part of this approach, several techniques were employed to analyze concerns against the workflow record as shown along with their comparative execution price. Chin and Gopal [3] focused on the information warehouse design, centered on a standard workflow metamodel that specifies the relation between workflow specification description and workflow performance features.

2.1 Mining Workflow

Data exploration in workflow records for uncovering different types of information about the workflow instances is discussed in several documents, such as [4–7]. Yifei and Lei [4] and George et al. [5] also recommend methods for instantly deriving the official design of an activity from a log of events related to the accomplishments of a procedure that is not supported by an activity control system. However, the work is restricted only to successive procedures. Deriving from the same kind of procedure records, Duan and Da Xu [8] obtained a workflow design that depends on Petri nets and integrates time details like small, maximal, and regular time placed at a certain level of the procedure.

At this moment, the details get connected to places of the produced Petri net-based workflow design. They also recommend a XML-based structure for storing and trading workflow records. Cuzzocrea [6] and Demurjian et al. [7] present an inductive learning component, used to support purchase and adaptation of successive procedure designs apart from generalization of execution records from different workflow instances to a workflow design while protecting all records. Moroa et al. [9] stated that since it handles workflow exemption and presents us a case study of procedure performance log. In fact, the objective of the contributors is to assist customers in handling exceptions once they occur.

2.2 Control and Manage Workflow

Several methods have been designed to keep track of the management of workflow performance. For example, Assunç~ao and Calheiros [10] focused on the case study of finalized review process details and indicated a device (called PISA) that brings together the study of the review pathway details and the focus on details obtained from the acting procedure resources of the company. They differentiate between three diverse views of procedure monitoring and controlling methods, other involved resources, and procedural issues. The PISA device contains data source connects that is responsible for the conclusion of access to surgical details resources.

3 Techniques Used in BI

In this section, different approaches and applications used in BI to provide efficient decision plans for the development of enterprise business in real time were discussed. Some of these data mining techniques, data presentation, report decision making, and statistical data analysis concepts are required to create effective business intelligence. The following sections contain a brief discussion on all these parameters.

3.1 Data Visualization

The primary function of BI is creating representation for analysis and evaluation of functions found in different data formats. Moreover, data visualization in BI consists of many ad hoc formats used for analytical process. Data visualization provides an effective connection to users as well as to the interactive database servers that are based on attributes in real-time synthetic data sets. It helps to analyze the relations attributes share with data items in patterns, trends, and exceptions of different dimensions for processing emerging data events. Finally, data visualization is more pervasive in BI as it represents data based on a graphical view of a company with semantic attribute learning and processing along with their corresponding relationships.

3.2 Data Mining Techniques

Data mining is defined as retrieval of relevant data based on attribute relations when processing of relational databases is done. It helps in the analysis of undefined relations based on object categorization in real-world systems. Based on its experience in some real-time business intelligence organizations, data mining is one of the analytic tools to analyze big data and data mining as well as perform user data analysis from different dimensions, finally classifying them based on identified relationships.

3.2.1 Data Classification

Fundamentally, classification is a process of grouping different parameters of attribute to enhance the support for taking the decisions in business intelligence. BI, gather data from various Web sites or openly accessible URLs for handling out the data on the fly. Classification identifies and processes information based on objects for easy identification in the real-time information stream assessment. Classification has many different techniques, for example, decision tree, support vector machines, and Naïve Bayesian. These are used for taking correct decisions and easy management of classifiers with well-organized capabilities and parameter processing.

3.2.2 Data Clustering

Clustering is an experimental information investigative method that finds out the superior answers for classification obstructions. Mainly, clustering is the job of assemblage a set of identical entities in a cluster or comparable cluster sets. The primary job of clustering is to identify appropriate information for statistical information investigation that is used in many information investigative areas, which com-

prises machine learning, pattern matching, information retrieval, image processing, and big data analysis.

3.2.3 Data Warehouse Algorithm

Data warehousing provides a distributed environment for managing relationships of different attributes. Distributed data warehousing algorithms are developed to process multiobjective data relations in relational data model.

3.3 Indexing Algorithm

This index algorithm extracts the required information from index formation with more accuracy. Index algorithm utilizes data set, and by indexing with clusters, this algorithm performs efficient data storage retrieval of record information. This method used for BI in store multimedia repositories with indexing and searching is based on keyword annotation and standard text-based indexing and data retrieval implementations that fetch and store data with different attributes.

4 Advanced BI Techniques in Different Real-Time Applications

Based on some real time experience and improve impact of information based problems to be contemporary in business oriented organizations. In this section, business intelligence research with different real-time applications like healthcare systems and telecommunication systems were discussed.

4.1 BI and Data Classification for Customer Relation Management (CRM)

In advanced BI, improve the complication of market and implement necessary operations like e-business implementations, knowledge implementation, and management and development of different plans in implementation of company's performance. In that, business intelligence is an indispensable; it facilitates the acquisition of data knowledge with factors to affect the increase of company operations. To develop modernly in company specifications, BI with customer-relationship management in those following operations is present.

4.1.1 Importance of CRM

In the present-day economy, a need of business exercises turns into a two-way correspondence among the organization and its clients. This correspondence depends on the benefits of the two parties: organizations that try to benefit, persist, and develop and clients who need to accomplish included esteem. The best organizations today are those that make their business forms in accordance with client desires.

4.1.2 Classification Approaches to Support for CRM

In customer-relationship management, feature selection is an effective attribute to improve company statistics with different relations. After attribute selection, classification is processed with the following classification approaches. They used default factors, and tenfold cross-validation is executed.

- (1) J48: The J48 method utilizes Quinlan's C4.5 calculation, which is an expansion of Quinlan's prior ID3 calculation. It creates a choice tree that can be utilized for grouping. J48 manufactures choice trees after an arrangement of named preparing information utilizing the idea of data pick up and entropy. Each characteristic of the information can be utilized to settle on a choice by part the information into little subsets.
- (2) Naïve Bayes: The Naïve Bayes classifier depends on Bayes hypothesis. Bayes hypothesis bargains a method for processing the back likelihood. It accepts that the impact of the estimation of an indicator on a agreed group is autonomous of the estimations of different indicators.
- (3) SVM: Support vector machine is a calculation that makes use of a nonlinear plotting to change the first preparing information into an advanced measurement. Inside this firsthand measurement, it looks through a choice limit which isolates the records of one class from the other class. The support vector machine discovers this choice limit utilizing bolster vectors and edges.
- (4) KNN: The closest neighbor classifiers contrast a specified test records and preparing records that are like it. The preparation records are depicted by 100 properties. Closeness is characterized by utilizing Euclidean separation.

4.2 *Business Intelligence and Analytics (BI&A) in Competitive Sports*

In this section, competitive sports presentation using business intelligence analytics with different attribute relations was discussed. From the last 10–20 years, business intelligence and analysis are an interesting concept in information systems with real-time developments.

Business intelligence and analytics (BI&A) rose as a vital subject in the train of information systems (IS). A portion of the creators characterizes BI&A as “the methods, innovations, frameworks, practices, strategies, and applications that break down basic business information to enable an undertaking to better comprehend its business and market and settle on auspicious choices.” As the definition proposes, the utilization of BI&A instruments by chiefs basically goes for exploiting the various wellsprings of accessible information and data to upgrade basic leadership inside associations.

4.3 Real-Time Intelligence for Adaptive Enterprise

This section describes the vision with real-time business intelligence for adaptive enterprise application of British telecommunication which contributes to the realization of real-time BI in different types of applications. BI tries to address these deficits by giving programming apparatuses that are altered for end business clients and convey business experiences continuously at the purpose of a choice. Current BI arrangements miss the mark regarding what is fancied.

They expect masters to run a measurable examination or an information mining procedure and set up reports that would then be able to be gotten to by business clients—they do not empower activities to be proliferated once more into business forms. BI items once in a while bolster client selectable information sources and constant information combination.

All present BI items rely upon preconstructed information stockrooms. There are no arrangements of ongoing business execution information accessible on the grounds that business action checking is still outside the BI space. Because of these breaks in the BI framework and manual mediations by experts, end business clients do not normally have constant access to data and cannot change forms continuously in view of bits of knowledge gotten from BI reports.

5 Comparison of BI Techniques

See Tables 1 and 2.

6 Scope of the Research

Due to the increasingly large amounts of data available in the server of the present-day organizations as well as the client expectations and high availability of data sources, it has become necessary for organizations to generate report across multitransactional systems of stored historical data. To address this problem in relational data assets of

Table 1 Comparison description of different data mining techniques used in BI

| Technique | Description (with the usage of BI) |
|---------------------|--|
| Data classification | In business intelligence, classification defines decision making between different outsourcing real-time processing systems |
| Data clustering | For effective data analysis in BI, clustering defines both statistical analysis with machine learning and information retrieval with different business intelligence applications |
| Data warehouse | To provide distributed management for BI and to define managing relations between different fields to store huge amount of relevant and irrelevant data to process multiobjective in data analysis of different BI organizations |
| Data indexing | To provide effective interface for BI based on required information for data storage and data management for BI to explore content-based indexing for multidimensional usage in different BI real-time applications |
| Data visualization | Identification of data in BI application and data visualization provides an efficient connection between users to access database server (where they were present in real-time processing) |

Table 2 Real-time application comparison description in BI

| Real-time approach | BI description |
|----------------------------------|--|
| Data warehouse healthcare system | Clinical business intelligence processes the procedure of classification for data analysis techniques. Data warehouse is necessary to develop healthcare system application |
| Classification for CRM | In trending implementation of trending advanced technologies that improve e-business implementations, knowledge implementation, and management and development of different plans in implementation of company's performance, BI is indispensable |
| BI for MA | This section describes the usage of mobile BI application using Native development of the platform with different relations in application development. A model mobile BI program is built to represent mobile OS sales information in a visible form to allow the customer to get a specific view of the information |
| BI&A | Business intelligence and analytics (BI&A) rose as a vital subject in the train of information systems (IS). A portion of the creators characterizes BI&A as "the methods, innovations, frameworks, practices, strategies, and applications that break down basic business information to enable an undertaking to better comprehend its business and market and settle on auspicious choices" |

data marts, cubes in data warehousing operations from a combination of different data sources are used. In Sects. 2 and 3, different algorithms and approaches designed for BI application development are discussed. To make efficient decisions for enterprise applications, BI offers various tools and techniques for extraction, transformation, and loading (ETL) in the construction of data warehouse, reporting, information retrieval, and querying with decision support systems.

To perform effective association, reporting, and statistical data presentation for effective decision making in enterprise applications, the implementation of a hybrid technique which consists of classification, association, and clustering for performing internal analysis and predictive analysis of high-dimensional business data is used. The hybrid data mining technique is also used to application design growth (AD growth) algorithm for effective data utility in transactional databases in business intelligence.

The details of the high utility itemsets are maintained in a tree-based details framework, known as utility design tree (UD tree), so that these utility itemsets can be generated efficiently with only two tests of data source in business intelligence. Experimental results show that the suggested methods, especially AD growth, not only decrease the number of applicants successfully but also outshine other methods considerably with regard to playback, especially when data source contains lots of long transactions.

The frequency of algorithms/techniques used was taken along the y-axis, and names of the techniques used are shown along x-axis in Fig. 1. From Fig. 1, it is evident that most of the researchers have used classification techniques rather than clustering and visualization techniques. The order of techniques used is the classification followed by cluster, visualization, BI&A, and indexing.

7 Conclusion and Future Work

This paper provides a brief literature study on the application of BI for relational data sources and discusses some related work on enterprises with a special focus on multirelational data sources. Various data mining techniques, approaches in data visualization in multidimensional data sources, reporting data representation, and statistical data analysis by using some data dissemination models were discussed. Indexing algorithms and calculations in relational data sources usage were discussed. The introduction of hybrid data mining approach that consists of data clustering, association, and classification in crystal report generation for transactional data sources was made. Application design growth calculations for mining high utility items from transaction data sources have been introduced. Attacker model usage for preserving privacy of outsourcing transactional data in business intelligence applications was found. It was observed from the survey that the decreasing order of BI techniques frequency is classification, clustering, visualization, BI&A, and indexing algorithms/techniques. It is concluded that classification, clustering, and visualization methods were more popular in the research groups.

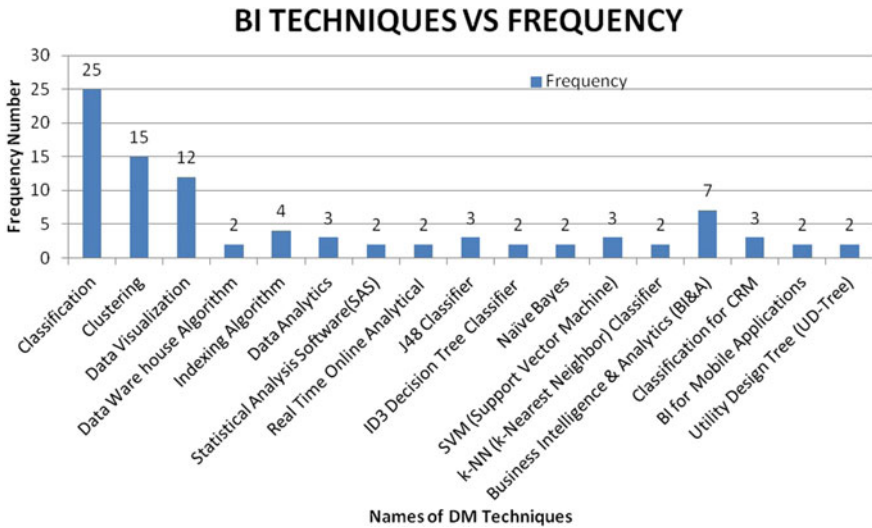


Fig. 1 BI techniques versus frequency number

It is required to apply new and more potential algorithms; namely, the hybrid algorithms and deep learning algorithm need to be applied for extracting decision support enterprise solutions and to make the entrepreneurs to carry out their business in an effective manner.

References

1. Gounder, M. S., Iyer, V. V., & Al Mazyad, A. (2016). A survey on business intelligence tools for university dashboard development. *Big Data and Smart City (ICBDSC), 2016 3rd MEC International Conference on IEEEXplore*, April 28, 2016.
2. Yoon, T. E., & Ghosh, B. (2014). User acceptance of business intelligence (BI) application: Technology, individual difference, social influence, and situational constraints. *System Sciences (HICSS), 2014 47th Hawaii International Conference on IEEEXplore*, March 10, 2014.
3. Chin, W. W., & Gopal, A. (1995). Adoption intention in Gss: Relative importance of beliefs. *ACM SIGMIS Database*, 26(2-3), 42-64.
4. Yifei, H., & Lei, H. (2015). The research of business intelligence system based on data mining. *Logistics, Informatics and Service Sciences (LISS), 2015 International Conference on IEEEXplore*, January 04, 2016.
5. George, J., Vijay Kumar, B., Santhosh Kumar, V. (2015). Data warehouse design considerations for a healthcare business intelligence system. In *Proceedings of the World Congress on Engineering 2015 Vol I, WCE 2015*, July 1-3, 2015. London, UK.
6. Cuzzocrea, A., Bellatreche, L., Song, I.-Y. (2013). Data warehousing and OLAP over big data: Current challenges and future research directions. In *Proceedings of 22nd ACM International Conference on Information and Knowledge Management (CIKM'13)* (pp. 67-70). San Francisco, CA, USA, October 27-November 01, 2013.

7. Demurjian, S. A., Blechner, M., & Saripalle, R. K. (2012). A proposed star schema and extraction process to enhance the collection of contextual & semantic information for clinical research data warehouses. In *Proceedings of the 2012 IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW)*, IEEE Computer Society (pp. 798–805). Washington, DC, USA, October 7–10, 2012.
8. Duan, L., & Li Da, X. (2012, August). Business intelligence for enterprise systems: A survey. Published in *IEEE Transactions on Industrial Informatics*, 8(3).
9. Moroa, S. M. C., Cortezb, P. A. R., & Rita, P. M. R. F. (2015). *Business intelligence in banking: A literature analysis*. Preprint submitted to Elsevier October 20, 2015.
10. Assunçãoa, M. D., & Calheiros, R. N. (2014). Big data computing and clouds: Trends and future directions. Preprint submitted to *Journal of Parallel and Distributed Computing*, August 25, 2014.

Performance Analysis of E-Governance Citizen-Centric Services Through E-Mitra in Rajasthan



Praveen Kumar Sharma and Vijay Singh Rathore

Abstract Implementing a citizen-centric approach to delivering government services is the need of the hour which creates and maintains a level of dialogue and trust between citizen and government. As per the study conducted so far, it was seen that implementing such system aids in increasing the public satisfaction and reduces cost. In technical terms, this process is called e-Governance. E-Governance is the application of information and communication technology (ICT), which acts as a medium for the delivery of government services, information exchange, communication transactions, integration of different systems and services between government and citizen (G2C) and government and government (G2G). In Indian scenario, National e-Governance Plan (NeGP) is an initiative taken by Indian government to provide all types of government services to the citizen of India. Rajasthan, a state of Indian country which is considered to be lacking behind in terms of technological advancement, is possibly one of the very few states where e-Governance initiatives are going on successfully from a very long time. Presently, there are many e-Governance practices prevailing in the state like Bhamashah Yojana, e-PDS, Rajasthan Payment Platform, eSanchar, iFact, Sampark, eBazaar, Raj Wi-Fi, Raj eVault, etc., but the nationally renowned e-Mitra launched in 2005 has a wider scope in terms of both people reach and government services offered. The research work presents an analysis of e-Governance citizen-centric services through e-Mitra in Rajasthan. The research work attempts to provide an insight into the ongoing e-Governance practice of e-Mitra to reveal how digitization can be used to enhance the reach and the actual effect of such government services.

Keywords Citizen-centric approach · e-Governance · e-Governance practices · ICT · e-Mitra

P. K. Sharma (✉)
Mewar University Chittorgarh, Chittorgarh, Rajasthan, India
e-mail: praveenvmou@gmail.com

V. S. Rathore
JECRC, Jaipur, Rajasthan, India
e-mail: vijaydiamond@gmail.com

1 Introduction

Citizens are the major factor in Indian democracy. E-Governance is a kind of information and new technologies provided by the government to the citizens. E-Governance services now become more reliable for the citizens day by day. At starting, e-Governance is only used in government sectors as Web portals [1]. Rajasthan government is not taking steps to be implemented in all the sectors as public, private, commercial, etc., so that it becomes more efficient to the citizens. In regarding the same, Rajasthan government implements a number of e-Mitra all over the states and this becomes the most progressive dynamism of Rajasthan government.

This e-Governance service provides clarity and facilities to proceed with something without difficulty at single place and closer to the citizens. The objective of the study is to know the kind of services provided by e-Mitra branches and also investigate the influence on the citizens due to the services. E-Mitra developed high-definition videos, voice clarity, and easy services in the field of education, health, and other private and public sectors. The main motive of e-Mitra is to offer a platform in rural area to pay bills for electricity, mobile as well as to fill examination form of different sectors [2].

2 Perception and Aim of E-Mitra

The objective of e-Mitra is to make a base that can permit government and private sector corporation to put in order their social and merchant goals for the profit of the village people in the distant corner of the state or country. The project of e-Mitra brings all the branches at single place which makes the system user-friendly as well as limpid. The objectives of e-Mitra project are described in Table 1:

Table 1 E-Mitra Project

| | | | | |
|-------------------------|--------------------------|-------------------------------|--|------------------------|
| Satisfies more citizens | No need of Agents | Helpful attitude of employees | Less time and effort to avail services | Ease of administration |
| Error-free transaction | Security of data | Good complaint handling | Adherence to citizen's charter | Paperless office |
| Less waiting time | Good location | Convenient time schedule | Lower cost to the citizen | Less corruption |
| Greater transparency | Equal opportunity to all | More predictable outcome | High clarity and simplicity of process | Single-window system |

- It will give problem-free advice to the peoples.
- Whole e-Mitra will reduce the different points to the peoples and remove the time loss.
- Magnified sources and such organization are not in the public sector services. In a regular manner, data exchanges in a non-building structure.
- The partnership project for staff deals with the public and private.
- Technological partner handles the office which not deals with the public, but it will be the government office.
- It will give the chances for jobs in all over the states.
- It will use the Internet, execution of series of jobs without manual intervention, and immediately provide data to the user [3].

3 Research Query/Structure for Project Study

This research pursues to recognize the issue analogous with assigning e-Governance from both employ approach and organization approach. The research attempts were provisional. The provisional questions of the research comprise the following:

- What are employee prediction from e-Mitra services and source of supply?
- What are user obstacles and issues confront using supply source and e-Mitra services?
- What influence easier and raise user knowledge with e-Mitra services?
- What are the key operators of the enlargement and accomplishment of e-Mitra services?
- To what expanse are people require and prediction involving in the blueprint and accomplishment of e-Mitra services?
- How are people recognizing prediction and impulse in e-Mitra services and resources includes in general blueprints and continues improvements of e-Mitra services?
- What requires do employee have in venture to recruit in e-Mitra services?
- Are there blueprints issues that simplify or proceed as obstacle to triumphant people e-Mitra synergy [4]?

4 Research Methodology

Research methodology hires for the study of qualitative as well as quantitative procedure. According to the objective of the whole study, it was suggested to managing the study for most of the e-Mitra branches of Kota (257 centres were elected) randomly from the total number of e-Mitra list of Kota. A more than 100 people's reviews were taken and studied. The data of two months were collected for the surveys and study.

Some question was managed for the pattern examinations which were furthered distributed in two parts:

1. Examination for e-Mitra business person and organization structures.
2. Examination for perception level, citizen's side view, and feedback procedure.

The aim of the project is to check the concept of the e-Mitra branches towards the technical and identified programmatic are within the division, department, laboratory, or area of focus. The technical explanation are:

- State being connected. A satisfactory connectivity's found to be at e-services and how the last KM connectivity will be got?
- E-services node follows the design of IT architecture? (Working process research included.)
- What will be benefaction of e-services in G2C and B2C process of transacting business?
- How the concession to design colour project protect at the centres?
- The accomplished importance of e-services being pursue?
- Are citizens' complaints are finding correctly?
- What is the procedure of courses to be provided for refreshers to the e-Mitra?

1. For the purpose of inventing the research on the examination was created in the form of an app. On the lead capture page of the app following description are as follows:

- About e-Mitra?
- Feature of e-Mitra?
- Services e-Mitra?
- Payment options?
- Help or query?

2. The body of opinion data inputs is touch by a set of printed text that are tested out quality of the services which is given to the peoples by the e-services/Mitra are:

- Understanding about the e-services.
- Attention given to the people who know about the services.
- The people travel a mile to reach e-services.
- Approved price given by the people to come at the e-services.
- Approx. time will taken by the people at e-services.
- No. of visiting which is made to the branches far getting a service.
- Comfort of getting running work.
- Services offer by the e-services are shown at the branches.
- The running day of convenience and beneficial.
- Pain bill is always correct.
- Total process required to progress a G2C case.
- Total days required to issue a G2C certificate.

- The working employers are need to be friendly and give attention to the customer.
- Satisfaction with the characteristics of the services.
- Average salary of the peoples utilizes the services of e-Mitra.
- Frequency of the people was for sum of the total services utilized the e-Mitra [5, 6].

5 Detection and Discourse

Appreciation/Motive: There was a great level of appreciation about e-Mitra approach of the 100 appellants, only 5, i.e. 5% of the appellants end user, have not witnessed of e-Mitra. On the other side, 85% of appellants were gently gathering favour from e-Mitra [7].

Approach of Advertising: Approach of advertising through which peoples appreciate about the benefits was several. Media is the best way to promote e-Governance.

Utilization design: Utilization of e-Governance services was steady for an enormous percentage of examine 1 community which is 75%. This giant proportion of employee progressively utilizes services at e-Mitra specify its marketability. This also guarantees steady of the representation in destiny time.

People stroll inside: Usually, 75 moves in all over day in at e-Mitra. A superior number of people moves in all over day are beneficial for both the entrepreneur and peoples. Some of the e-Mitra branches delineate 150 deals in a day.

People price of service: The approx. mile which is travel to the people to come at e-Governance was approx. 2 km. The change in rural area is more but all the people who are come to the e-Governance say there is no problem to coming by covered such a mile. The average money which paid by the people for coming at the e-Governance is approx. 15 Rs. Minimum time required to come at the e-Governance are 20 min only approx. one visiting is required to getting a service from e-Governance.

Contentment status: 50% of the people who getting the services are satisfied to working of the employ. Only 10% people are complaining for not getting a proper facility.

- Average time or the facility given to the customer are lower to the model (50% are satisfied).
- The area of branches of e-Governance are good (70% are satisfied).
- Only one window has different type of services.

Service list provided to visitors: Only 30% of e-Governance centres will show the service list of the e-Governance and fee which have provided the secure service equation. So, in the service list is important part of the promoting the clarity or fraud-free services. The e-Governance branches were the service lists shown to the customer; there may be the different charges are taken by e-Governance to the people. Here the proper menu or service bill will make to the people trust on the e-Governance services [8].

Building an atmosphere of the place: The building atmospheres of the place are generally suitable or sufficient. The location will always be checked by the experience people. Clarity of the place is satisfying 55%. People wait time will be rated as low by 35% as average by 20%. Average service efficient status is 65%.

Facility eminence and perfection: The rightness status and pay bill are correct reported by 68% of people. This will give the trust of the people in the PPP model of e-Mitra. The major problem in e-Governance is to connect the e-Governance counters.

Input progress and the time period for G2C certificate: Total progress inputs are requiring to progress the G2C case which was a approx. three documents. Approx. sometimes as high 12 days and 4 follow up visits to visit C2C certificate variable. Here the G2C paper needs the signature of district officer due to which delay in the issue is there.

Experience officers handle e-Governance: Here we will agree that the officers are friendlier and give personal attention to the people which come at the e-Mitra. The e-Mitra will be improved the status of the governance due to such an officer which give their personal attention and help as much as possible. Due to the good service are given to the people, the business is run good which shown by the e-Governance.

Satisfied status from e-Governance model: There is status of satisfaction which is high with the e-Governance model in the people and user. There is 60% of the total people report their comfort with the service. The citizens who use service of the e-Governance are highly satisfied without feeling inconvenience.

Salary and qualification status of people use e-Mitra: Approx. salary of the people utilizes the services of e-Mitra was put in the medium salary group. The government is more useful for those users which come in the medium salary. A large amount of the people in society (35%) utilize service of the e-Mitra are not have common education degree.

People involve in the service pattern and back support process: Here the government does not involve in the people development e-Governance service. There is very low approach of people to give suggestion to development of the service. Here no feedback is for user to complain or give suggestion for development of the services. Government team do not involve properly. Language is also a great problem to the people involved in improvement of the services. Even Hindi was also not known by most of the citizens [9, 10].

6 Conversation and Proposal for Augmentation

Pleasing people in e-Governance: Rajasthan government requires to pleasing end users prosperous in e-Government execution. This needs span of frequentative and unsegregated arranged step and sketch procedure such as follow:

- Managing a data.
- Technology requires evaluation.

- Regulate the accessibility of relevant data and favour to encounter user requirement.
- Intelligibility examines.

Those are dominant for extension and nourishment of e-Government attempts [11, 12].

Capture e-Mitra speculators and end users: Numerous types of difficulty can be recognized and rectified by pleasing users in the genuine blueprint. There is an extent of ability through which e-Mitra speculators were attracting such as:

- Pivot category and interrogation.
- Motivate real-time remark and proposals.
- Record folder and agreement record investigation.
- Support influencing each other online [13].

Site resolution for e-Mitra work surface: At least one e-Mitra must be unlocked inside span of 5 km area in cities. In villages, e-Mitra should be set up according to the number of people overspread.

7 Conclusion

Through the e-Governance services, the civic services can be implemented easily. The criteria of e-Governance awareness are depending on the age of citizen, education, occupation, and Internet literacy. Through the help of the e-Governance services, a citizen can interact directly with government and government policies.

Due to this feature, it creates a transparency between citizens and government which will help in development of nation. Citizens can benefit civic services provide by e-Governance at anytime and anywhere in nation according to their needs. Truly the housewives are not fully aware about the great significance and implementation of e-Governance. Due to e-Governance services, citizens can avail civic services directly without taking help of intermediate. The governance system removes the queue system; due to this system citizen can avail its feature easily and fast.

E-Governance services can be accessed by citizens at anytime and anywhere to interact with government. Citizens also like this system. Citizens also feel that implementation of e-Governance system helps them very efficiently. Due to these features of e-Governance, all works can be done easily and fast without standing in queue. This system reduces the travel cost and efforts of citizens.

Due to transparency, clarity features of e-Governance services help employee to satisfy more citizens. This will get increment and service quality. The manual system cost is very high as compared with e-Governance services this reducing the service cost. Frauds of "Birth and Death Certificates" can be reducing by help of e-Governance system. Due to reducing of travelling cost and service cost, the e-Governance system saves approx. 250 Rs per year. In near future, all the civic services can be handled on the e-Governance platform which is very hard to be recovered by the manual method.

References

1. Foley, K. (2006). Using the value measuring methodology to evaluate government initiatives. In *Proceedings of the 2006 Crystal Ball User Conference*, May 1–3. Denver. Colorado.
2. Freed, L. (2009). *E-Government satisfaction index*. ForeSee Results periodical publication.
3. Singhal, R. (2013). An assessment of benefits delivered to citizens through e-governance from e-Mitra in Jaipur. *International Journal of Computational Science, Engineering & Technology*, 1(1), 43–51.
4. Heeks, R. (2006). *Understanding and measuring e-Government: International benchmarking studies*. Paper prepared for UNDESA workshop on “e-Participation and e-Government: Understanding the Present and Creating the Future”, July, 27–28, Budapest, Hungary.
5. Akman, I., Yazici, A., Mishra, A., & Arifoglu, A. (2005). E-Government: A global view and an empirical evaluation of some attributes of citizens. *Government Information Quarterly*, 22, 239–257.
6. Tejseev, S., Sarangdevot, S. S. (2014). Integration of ICT and e-governance in Rajasthan. *Indian Journal of Computer Science and Engineering*, 1(2), 177–183.
7. <http://doitc.rajasthan.gov.in>.
8. Kunstelj, M., & Vintar, M. (2004). Evaluating the progress of e-Government development: A critical analysis. *Information Policy*, 9(3–4), 131–148.
9. Bhatanagar, Subhash. (2009). *Unlocking e-Government potential—Concept, cases and practical insights*. New Delhi: Sage Publications.
10. Rajasthan state Web portal for e-evaluate Citizen centric service. <https://rajasthan.gov.in>
11. Gupta, P., & Bagga, R. K. (2008). *Compendium of e-Governance—Initiatives in India*. University Press.
12. <http://emitra.rajasthan.gov.in>.
13. Heeks, R., & Bailur, S. (2007). Analyzing e-Government research: Perspectives, philosophies, theories, methods, and practice. *Government Information Quarterly*, 24(2), 243–265.
14. <https://rajasthan.gov.in>.

A Critical Study on Disaster Management and Role of ICT in Minimizing Its Impact



Pratibha Choudhary and Rohit Vyas

Abstract Disaster means “Bad Star” in Latin and is defined as an impact due to a natural or man-made hazard that results in the huge casualties and damage to resources. Recent events and studies have shown that disaster preparedness is no longer considered as a choice and has become a mandatory process irrespective of the geographical area and its distribution. The types of risk vary and increase depending on the geographical location of a country. If they are targeted proactively, the consequences of natural and man-made disasters and the vulnerabilities to which people are exposed can be mitigated. It has been proven from past experience and practice that the damage caused by any disaster can be minimized largely through careful planning, mitigation, and prompt action. In disaster prevention, mitigation, and management (disaster management), information and communications technology (ICT) can play a key role. Different available technologies, including telecommunication satellites, radar, telemetry, and meteorology, allow remote sensing for early warning. ICT includes both traditional media (radio, television) and new media (cellular broadcasting, Internet, satellite radio), which can play an important role in educating and raising public awareness of the risks of potential disaster. This research work highlights the issues of disaster management in relation to the Indian subcontinent. It explores the role of “National Disaster Management Authority” (NDMA) established in 2005 by the Government of India to study and minimize the effect of a disaster using various ICT tools and techniques.

Keywords Disaster · Disaster management · ICT tools · ICT techniques · NDMA

P. Choudhary (✉) · R. Vyas
Government College of Engineering & Technology, Bikaner, India
e-mail: pratibhacivil.gcet@gmail.com

R. Vyas
e-mail: rohityas@protonmail.com

© Springer Nature Singapore Pte Ltd. 2020
M. Pant et al. (eds.), *Performance Management of Integrated Systems and its Applications in Software Engineering*, Asset Analytics,
https://doi.org/10.1007/978-981-13-8253-6_18

183

1 Introduction [1, 2]

Disasters are as old as human civilization, but from some of the past decades, damage caused by them focused the problem in national and international concern. There are number of natural and man-made disasters happened vigorously. Between 1994 and 1998, the United Nations published a report that an average of 428 disasters per year, but from 1999 to 2003, this figure increased from an average of 707 disaster events per year, which was approximately 60% in previous years (Fig. 1).

The greatest increase was in countries with poor human growth and civilization, which experienced a 142% increase. The figure shows the decade's deadliest disasters (1992–2001). Drought and famine were part of the world's deadliest disasters, followed by floods, technological disasters, earthquakes, windstorms, extreme temperatures, and others.

2 Condition of Disaster in India [3, 4]

In India, since last few centuries, a number of disastrous events happened like the Bengal Famine (1770) in which almost 1 crore casualties were reported, Calcutta cyclone (1737), earthquake in Gujarat (2011), Tsunami in Kerala, Cyclone in Andhra Pradesh, floods in Bihar and landslides in Kedarnath. There are some other man-made disasters like 26/11 terrorist attacks; Bhopal gas tragedy and many more are responsible for the disaster.

It was not only affecting the human life but also reducing the chances of growth of any country. Government of India passed disaster act in 2005 and established the department by name National Disaster Management Authority (NDMA) in the same year. The basic aim of this authority is to aware the people about modes of disaster and some of the guidelines and mitigation techniques to minimize the effect of disaster.

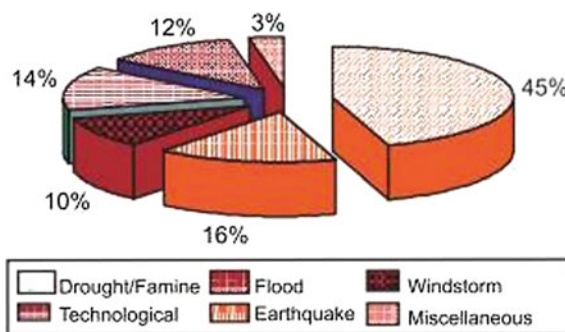


Fig. 1 Fatalities due to disaster

3 Need of Disaster Management [4, 5]

Disaster management includes the sum total of all activities which cause disaster, its measure, and guidelines before happening of disaster and most importantly action taken during and after a disaster with the purpose to avoid it due to which we are capable to reduce its impact or recover from its losses. Disaster management classified disasters according to their mode of occurrence and its mitigation techniques which are as follows (Table 1 and Fig. 2).

Table 1 List of natural and manmade disasters

| S. No. | Type | Disaster |
|--------|---------------------------------------|--|
| 1. | Geological disaster | Earthquake, landslide, tsunami, dam burst, volcanic eruption, mine fire etc. |
| 2. | Water and climatic disaster | Tropical cyclone, cloudburst, tornado and hurricane, landslide, floods, heat and cold wave, drought, Snow avalanche, hailstorm, sea erosion |
| 3. | Environmental and biological disaster | Environmental pollutions, desertification, deforestation, pest infection, human/animal epidemics, food poisoning, pest attacks, weapons of mass destruction (nuclear and hydrogen) |
| 4. | Chemical and industrial | Chemical, fire, gas leakage, oil spilling etc. |
| 5. | Accident related | Road accidents, building collapse accidents/air crash, electric accidents, rural/urban fire, bomb/serial bomb blasts, mine flooding, forest fires etc. |



Fig. 2 Disaster management cycle by NDMA [3]

4 Mitigation Techniques [6, 7]

The most important part of mitigation is the full understanding of the nature of the risk. The types of threat facing each country with different regions are different. Some countries are prone to flooding, others have experienced tropical storm damage, and it is known that some regions are in the earthquake zone. Most countries face various threats, and all face the possibility of technological disasters caused by industrial development.

The effects of threats are likely to occur, and the harm they are likely to cause depends on the situation in the region: the people, their homes, daily life sources, and infrastructure. It is different from each country. Knowing the types of hazards that may be encountered is essential for any particular location or country. Mitigation techniques are the methodologies used to reduce the effect of the disaster and its conditions to reduce the scale of a future disaster.

Activities to mitigate the risk caused by the disaster itself or the elements exposed to the threat can therefore be focused. Examples of disaster-specific mitigation measures include water management in areas prone to drought, removing people from hazardous areas and strengthening structures to reduce damage in the event of a hazard. Some of the mitigation techniques are as follows. There is some mitigation techniques which can be useful during disaster are as follows.

Floods: Flooding is India's number one natural disaster, and climate change can lead to inland, coastal areas, and all parts of the country. The following flood mitigation techniques are used.

- If you live in a high-flood risk area, raise your furnace, water heater, and electric panel in your home.
- Install “check valves” to prevent flooding into your home drains.
- Homeowners can build barriers (such as sandbagging) to prevent floodwater from entering your home when practical.
- Seal walls with waterproofing compounds in your basement.

Droughts: Drought is the most popular natural hazard; after drought, it evolves over months or even years. It may affect a large region and causes little structural damage. The impacts of drought can be reduced through preparedness and mitigation which are as follows:

- By installing artificial water sources like dam, canals, etc.
- By using contour bunds, trenches, and stone walls.
- By raising low-water crops, etc.

Cyclones: Cyclones are not so common in India due to the position of Himalaya. But in the past, there are major cyclones due to which crores of life were lost. Some of the mitigation techniques are as follows:

- By installing cyclone shelters.
- By using rigid engineering building.
- Trees like pine which has conical leaves are best suited to reduce cyclone.

Earthquake: It is an uncontrolled phenomenon, but we can minimize its effect by using the following techniques:

- By using lightweighted building.
- By installing dynamic foundation below buildings.
- By using shear walls in buildings.
- By using heavy plantations which are shock absorbers.

Some miscellaneous mitigation techniques used in disaster management.

- By using new technologies like GIS, GPS, and other networking methods for predicting the disaster.
- By developing the cities and other areas which are capable to handle disasters.
- By avoiding the conditions of riots/wars.
- By developing safe road/rail network to minimize the accidents, etc.

5 Conclusion

It is concluded that the occurrence of disaster is not controllable, but by using some of the techniques, its effects are minimized. It has been proven from past experience and practice that the damage caused by any disaster can be minimized largely through careful planning, mitigation, and prompt action. In disaster prevention, mitigation, and management (disaster management), information and communications technology (ICT) can play a key role. Remote sensing for early warning is made possible by various available technologies.

References

1. Sarkar, S., & Sarma, A. (2006). Disaster management act, 2005: A disaster in waiting? *Economic and Political Weekly*, 3760–3763.
2. Alzaga, A., Varon, J., & Nanlohy, S. (2005). Natural catastrophes: Disaster management and implications for the acute care practitioner. *Critical Care and Shock*, 8(1), 1–5.
3. Erramilli, B. P. (2008). Disaster management in India: Analysis of factors impacting capacity building.
4. Rao, K. H., & Rao, P. S. S. (2008). Disaster management. Serials Publications.
5. Anderson, J., & Bausch, C. (2006). Climate change and natural disasters: Scientific evidence of a possible relation between recent natural disasters and climate change. Policy Brief for the EP Environment Committee. IP/A/ENVI/FWC/2005-35.
6. Burton, I. (2002). Risk management and burden sharing in climate change adaptation and natural disaster mitigation. Discussion paper for UNDP expert group meeting on integrating disaster reduction and adaptation to climate change, Cuba, June 17–19, 2002.
7. Klein, R. J. T., Schipper, E. L., & Dessai, C. (2003). Integrating mitigation and adaptation into climate and development policy: Three research questions. Working paper 40, Tyndall Centre for Climate Change Research, UK.

Development of Arduino-Based Compact Heart Pulse and Body Temperature Monitoring Embedded System for Better Performance



Sandeep Gupta, Akash Talwariya and Pushpendra Singh

Abstract With lifestyle changes in modern time, ease of living due to technological advancement, and increase in urbanization and globalization, there is an increase in the cases of humans suffering from a vast variety of harmful diseases. According to the fundamental principle of protection from these harm diseases, two parameters of human body i.e., the current status of body temperature and running heartbeat measurement on regular basis, are vital and essential. With the advent of ICT tools and many advanced medical devices, these activities can be recorded with user-friendly display and interface in real time which can prove to be of more value and use when there is no nearby facility of hospital and medical care. This paper can create awareness about the one's actual severity of sickness. The aim of this research work is to present medical devices which are portable and compact in size and can be easy operated without expertise for measuring and showing the body temperature and running heart pulse. In this research work, a processing assistive integrated heart rate with body temperature embedded supervising device is developed. The system provides the information of heart rate via serial communication on the PC or laptop and body temperature on liquid-crystal display (LCD). This system is very useful to monitor conditions at remote places. The proposed device includes the various applications such as Arduino Uno microcontroller system, various sensors, transmission system, and interfacing. The proposed system is economical, has compact design, and is a lightweight instrument. This system has been tested using data sets, and based on the outcomes of this device, it is concluded that it gives comparatively better performance than old hand measuring system.

S. Gupta (✉)

Electrical Engineering Department, JECRC University, Jaipur, India
e-mail: jecsandeep@gmail.com

A. Talwariya · P. Singh

Electrical Engineering, JKLU, Jaipur, Rajasthan, India
e-mail: akash.talwariya@gmail.com

P. Singh

e-mail: pushpendrasingh@jklu.edu.in

Keywords Arduino uno · Body temperature · Heart rate · Pulse amplification · Pulse detection

1 Introduction

The heart rate can be simply termed as the sound of heart. The cardiovascular system can be assessed by heart rate [1]. The work of human heart is to pump the oxygen-rich blood to the muscles. It also carries cell waste away from tissue and cells. Heart rate can never be considered as the constant; it always varies according to the demand of muscles for absorbing oxygen and exerting carbon dioxide which mostly occur during sleep or exercise [1, 2]. Normally, the heart rate of the resting person is about 70 bpm for adult males, and for females, it is about 75 bpm. The heart rate monitoring system is the simple device which takes the instantaneous sample of one's heartbeat and calculates the heartbeat per minute, and the calculations can be easily monitor the current condition of heart.

On the other hand, body temperature is also the factor which generally indicates the body condition. Normally, the human body temperature is neighboring to $98.6\text{ }^{\circ}\text{F} \pm 0.7\text{ }^{\circ}\text{F}$, and it varies on the basis of measurement geographical place and the exertion of person [3]. The temperature of the body is measured through skin heat. When a person does some activity, the blood vessels explicate to bear the excess amount of heat on surface of body and sweating occurs. Then, the sweat dehumidifies and this process helps to regulate the body temperature and make body cool. When the human is cold, the blood vessels become narrow and the blood flow gets reduced to sustain body heat. In that situation, the human may stealers shivering appear involuntary and irregular muscle construction of the body.

Timely measurements of the human body temperature and heartbeat rate are important of superior treatment. In the current scenario, the environment is highly polluted and the concern regarding health is the priority for human beings. Nowadays, people are aware and spend a lot of money for their health, but unfortunately, in most of the cases, we found that the medical treatment is highly costly and the person does not able to take the advantage of technology on time. Therefore, the unexpected incident occurs due to the delay of treatment. As the heart rate and the body temperature are the most vital notable indices of human health, an affordable device will be very helpful to measure these parameters. In general, these types of equipments are very bulky and costly. Moreover, some devices [4–8] are available in the market that can provide raw medical calculations of data for patients and doctors, but the patients cannot able to interpret medical measures into the meaningful diagnosis due to lack of medical background of the literacy of the humans. Therefore, an autonomous and economical system for continuous heart rate measurement and temperature monitoring systems are very much essential.

In this research paper, a processing assistive integrated heart rate with body temperature embedded supervising device is developed. Since the system provides both the information of heart rate via serial communication on the PC or laptop and body temperature on liquid-crystal display (LCD), this system is very useful to monitor conditions at remote places. The rest of the paper shows the existing approaches, our system model, and system output performance.

2 Existing Approaches

Nowadays, optical and electrical methods are used for heart rate measurement. Electrical methods required such heavy or bulky straps around the chests, but the optical method does not require these kinds of straps; the results are more effective and efficient [9]. Optical technology is cheap in cost and uses the powerful LED and light-dependent resistor (LDR) to sense the pulses [7]. The signal is amplified by the amplifying circuit, and then, signal gets filtered with band-pass filter. The filtered and amplified signal is then sent to microcontroller (in our case, it is Arduino Uno R3). Then, the microcontroller checks the validity of analog signal and compares the signal with standard voltage. The microcontroller counts the great pulse and displays it on LCD. There are some more approaches that are exist such as Tx and Rx; the results of both these approaches are found incorrect in some cases because the signal varies for different persons and the calibration of these devices is difficult.

The objective of this study is to develop a device which monitors the heart rate and temperature at affordable price; it should be accurate, durable, and user-friendly; to display the results, the device is able to connect with PC, laptop, and LCD.

3 Heart Pulse and Body Temperature Monitoring System by Arduino Uno with Serial Data Communication

To construct an affordable and efficient device to measure the heart pulse and body temperature, an infrared Tx and Rx technique has been used to measure the pulse by calculating the change in blood flow throughout the fingers. For the measurement of body temperature, thermostat is connected to device as a sensor. Two microcontrollers are connected in this device. The first one is program in manner to count the pulse rate, and the second one is program to calculate the human body temperature. These both sensors are connected on Arduino Uno board. The board shows the results on LCD through serial cable. Figure 1 shows the block diagram of pulse controller. Figure 2 shows the block diagram of body temperature monitoring system.

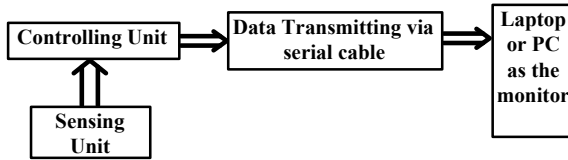


Fig. 1 Block diagram of the heart pulse monitoring system

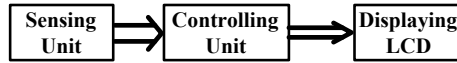


Fig. 2 Block diagram of body temperature monitoring system

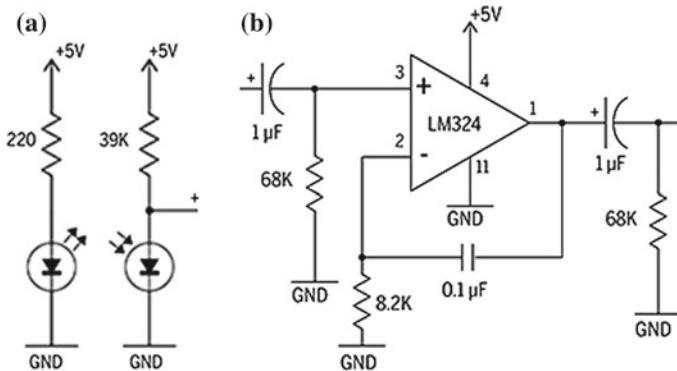


Fig. 3 **a** Pulse detection circuit; **b** schematic diagram of RC high-pass filter (HPF)

3.1 Heart Rate Monitoring System

To measure the volume of blood in the blood vessels, we have used the optical sensor to sense the heart pulse. This pulse sensor consists of the IRTx (IR transmitter) and IRRx (IR receiver). Tx transmits the infrared light to the fingers, and due to the blood flowing in blood vessels, it reflects back to the Rx. This amplitude of light reflection is converted into the pulse. In this paper, pulse sensor is used and this pulse sensor consists of three pins Vcc, GND, and Va as shown in Fig. 3a [8, 10], where Vcc is connected to +5 V of Arduino, GND is connected to the ground of Arduino, and Va is connected to the A0 pin of Arduino. If we connect the sensor to the Arduino, then the green-colored LED glows to the sensor and it makes sensor activated.

The waveform of the signal is synchronized with the heartbeats, but the signal is so weak and consists of much noise. Therefore, this weak signal is firstly passed to the OP-AMP LM324 and gets the DC component. Then, the signal is passed to a RC high-pass filter (HPF) as shown in Fig. 3b [11]. After that, the signal is passed through low-pass filter (LPF) which consists of an OP-AMP circuit. Figure 4b shows the further filtering and amplification of the output signal. The two stages of filtering

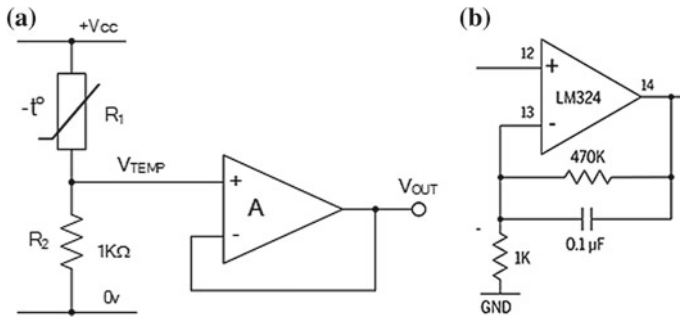


Fig. 4 **a** Schematic circuit diagram of NTC 10 k thermistor; **b** filtering and amplification of the output signal

and amplification convert the input signal to near transistor–transistor logic (TTL) pulses which are synchronized with the heartbeat.

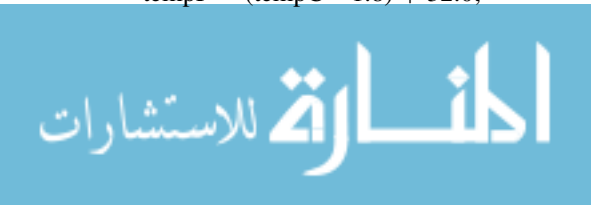
If the heartbeat is detected that the LED connected at the output of the second stage of the signal continuously blinks, at the final stage a simple non-inverting buffer is implemented so that the output impedance will be lower. After that, the output pin is connected to the analog pin A₀ of the Arduino Uno.

3.2 Body Temperature Sensing System

NTC 10 k thermistor is used for measuring body temperature as shown in Fig. 4a. The sensor is the accurate to measure body temperature [12]. Vcc is connected to +5 V terminal, the second terminal of the thermistor is connected to the OP-AMP, and the output of the OP-AMP is connected to A₀ pin of Arduino Uno. The last pin is connected to the ground pin of Arduino Uno.

For body temperature sensing from device, the following codes are used:

```
//read the temp
vo = analogRead(TR_PIN);
vo = vo/(1023.0 /5.0);
//voltage divider calculation
//vo = 5 * r2 /(r1 + r2)
//solve for r2
//get the exact value for voltage divider r1
r2 = (vo * r1) /(5.0 - vo);
//equation from data sheet
tempK = 1.0 /(a1 + (b1*(log(r2 /10000.0))) + (c1 * pow (log(r2 /10000.0), 2.0)) +
(d1 * pow(log(r2 /10000.0), 3.0)));
tempC = (tempK - 273.15);
tempF = (tempC * 1.8) + 32.0;
```



The above code is partial coding not a complete coding; this is only for understanding the concept.

3.3 Heart Pulse Data Processing Through Arduino Uno to Processing App

For data processing, we will send data from a single sensor to the program on a PC or laptop. The program written in the Processing app will graph the output of the sensor on screen. This is the way to find out a sensor's output corresponding to the physical events that it senses. Asynchronous serial communication, which we will see in this paper, is one of the most common means of communication between a microcontroller and another computer. The processing sketch in this work graphs the incoming bytes. Graphing a sensor's value like this is a useful way to get a sense of its behavior.

4 System Overview

In this research work, we have used two Arduino Uno programming boards with the ATmega328PU microcontroller. First of all, we have coded the microcontroller and interfaced the liquid-crystal display (LCD), thermistor, heart pulse sensor, and the Processing app. Figure 5 shows the internal connections of the proposed device. Firstly, the heart pulse sensor fetch the data from the finger then the microcontroller and calculate the threshold value then store the Bpm (Beats per minute) and then transfer this Bpm data to the processing app and this processing app helps us to monitor the data by the continuous pulse system and the Bpm displayed at the corner of that app as shown in Fig. 6. The second board interfaced with thermistor and LCD calculates the temperature in Kelvin (by Steinhart–Hart equation mentioned above), then converts it to the Celsius and Fahrenheit, and displays directly to LCD as shown in Fig. 7.

4.1 Performance of the Monitoring System

In this paper, we have tried to implement a low-cost heart rate and body temperature monitoring device using the pulse sensor and thermistor. In this system, it is possible to measure the heart rate and temperature of the patient continuously and can be treated in time for serious illness aspect. Actual versus measured heart rate data for 20 min are shown in Fig. 8. Actual versus measured body temperature data for 20 min are shown in Fig. 9. After the data analysis, it can be say that this monitoring device

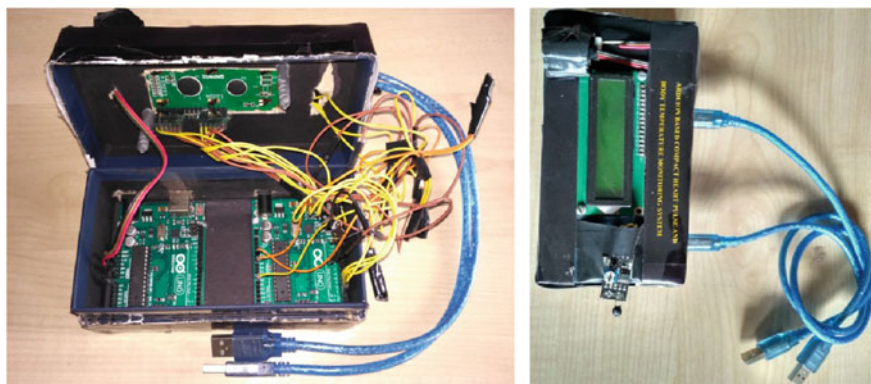


Fig. 5 Internal connections of Arduino-based monitoring device



Fig. 6 Pulse output of the monitoring device



Fig. 7 Temperature displaying on LCD

Fig. 8 Actual versus measured heart rate data for 20 min

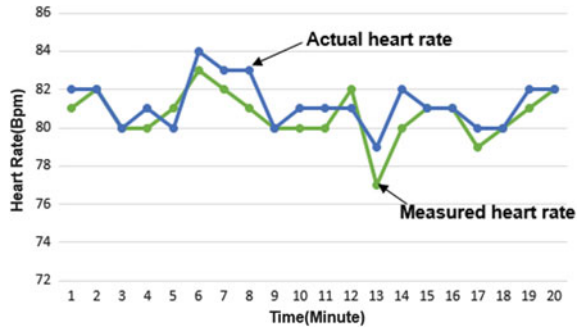
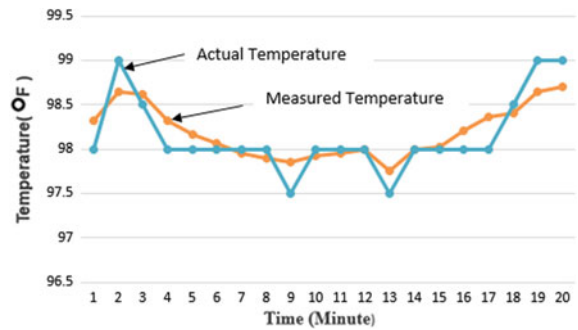


Fig. 9 Actual versus measured body temperature data for 20 min



provides much support to know the present condition of the patient if there is no doctor or clinic nearby.

5 Conclusion

This paper has designed a heart rate and body temperature monitoring embedded system. The final outcome of this paper approach is a portable heart rate and body temperature measurement system with a suitable architecture which can be applicable in medical and home appliances for health monitoring system of patients. This system has been tested for some valid signals such as heart rate and body temperature. And from the outcomes of this device, it has given comparatively better performance than old hand measuring system. Furthermore, this device can be improved with the help of different parameters such as retinal size, blood pressure, weight, and age which can be included to control the parameters in the future.

References

1. Kin, G., Paul, H. (2016). Medical instruments and devices: Principles and practices. *IEEE Pulse*, 7(2) (Books reviews).
2. Khandpur, R. S. (2003). *Handbook of bio-medical instrumentation* (16th ed.). Tata McGraw Hill publishing Co Ltd.
3. Celler, B. (1995). Remote monitoring of health status of the elderly at home. *International Journal of Biomedical Computing*, 40(2), 147–153.
4. Landreani, F., et al. (2016). Beat-to-beat heart rate detection by smartphone accelerometers. *European Heart Journal*, 37 (England: Oxford University Press).
5. Gogate, U., Marathe, M., Mourya, J., & Mohan, N. (2017, April). Android based health monitoring system for cardiac patients. *International Research Journal of Engineering and Technology*, 04(04), 1628–1634.
6. Hanna, K. J., & Hoyos, H. T. (2017, April 18). Compact biometric acquisition system and method. U.S. Patent No. 9,626,562.
7. Chatterjee, S., et al. (2016). Microcontroller based automated life savior—Medisûr. In *Proceedings of the International Conference on Computational Science and Engineering*. CRC Press.
8. Takemura, T. (1983, September 6). *Pulse detection circuit*. U.S. Patent No. 4, 403,193.
9. Yazicioglu, R. F., et al. (2016). Low-Power Biomedical Interfaces. In *Efficient sensor interfaces, advanced amplifiers and low power RF systems* (pp. 81–101). Springer International Publishing.
10. Chu, C.-T., et al. (2017). Non-invasive optical heart rate monitor base on one chip integration microcontroller solution. In *6th International Symposium on IEEE Next Generation Electronics (ISNE)* (pp. 1–4).
11. Gupta, S., & Tripathi, R. K. (2013). A pole placement controller for CSC based STATCOM with genetic algorithm. In *3rd IEEE International Advance Computing Conference (IACC)* (pp. 931–936).
12. George, B., Roy, J. K., Kumar, V. J., & Mukhopadhyay, S. C. (Eds.) (2017). *Advanced interfacing techniques for sensors: Measurement circuits and systems for intelligent sensors* (Vol. 25). ISBN: 978-3-319-55368-9.

Performance Evaluation of Learners for Analyzing the Hotel Customer Sentiments Based on Text Reviews



Dilip Singh Sisodia, Saragadam Nikhil, Gundu Sai Kiran and Hari Shrawgi

Abstract The world is in the midst of a digital revolution, and thus, it is natural that businesses are leveraging technology to position themselves well in the digital market. Travel planning and hotel bookings have become significant commercial applications. In recent years, there has been a rapid growth in online review sites and discussion forums where the critical characteristics of a customer review are drawn from their overall opinion/sentiments. Customer reviews play a significant role in a hotel's persona which directly affects its valuation. This research work is intended to address the problem of analyzing the inundation of opinions and reviews of hotel services publicly available over the Web. Availability of large datasets containing such texts has allowed us to automate the task of sentiment profiling and opinion mining. In this study, over 800 hotel reviews are collected from travel information and review aggregator site like Trip Advisor, and after pre-processing of collected raw text reviews, various features are extracted using unigram, bigram, and trigram methods. The labeled feature vectors are used to train binary classifiers. The results are compared and contrasted among ensemble classifiers, support vector machines, and linear models using performance measures such as accuracy, F-measure, precision, and recall.

Keywords Sentiment analysis · Opinion mining · Customer reviews · Binary classifiers · Ensemble classifiers · Naïve Bayes classifiers · Support vector machines

D. S. Sisodia (✉) · S. Nikhil · G. S. Kiran · H. Shrawgi
National Institute of Technology Raipur, Raipur, India
e-mail: dssisodia.cs@nitrr.ac.in

S. Nikhil
e-mail: nikhilnvm96@gmail.com

G. S. Kiran
e-mail: kiransujji1349@gmail.com

H. Shrawgi
e-mail: shrawgi.hari@gmail.com

1 Introduction

The world is in the midst of a digital revolution, and thus, it is natural that businesses are leveraging technology to position themselves well in the digital market. Travel planning and hotel bookings have become significant commercial applications. In recent past, there has been a tremendous growth in online discussion forums and review aggregator sites, for example, Trip Advisor [1, 2], in which the essential characteristics of a certain review are the associated sentiment or opinion. Sharp insights can be captured from the feelings of customer reviews on such platforms [3]. Opinions are cast into either positive or negative based on the words contained in the text of the review. If, for example, there are a lot of words having positive connotations, then it will be flagged as positive, otherwise as negative [4].

One of the other inputs from customers is the star ratings provided by them. At first sight, this seems to be a useful measure, but it is not so. Ratings are not expressive of the experience a customer has. Most of these are meaningless; moreover, a lot of these reviews belong within 3.5–4.5 ratings. A better approach is to turn words into quantitative measurements and use these measures to attribute connotations to reviews.

This study aims to improve this approach by including a feature of sentiment analysis that can classify reviews into positive or negative while maintaining accuracy [5]. Polarities associated with the reviews are also determined that help in categorizing it using semantic orientations [6]. Reviews are then assigned to a positive class or negative class, considering the overall semantic orientation with respect to the phrases extracted from them [7]. The classification model is trained on these extracted orientations, and thus, the predictions are made by the model [8].

Thus, a complete automation of the sentiment analysis procedure is achieved, starting from extraction, analysis, and classification of data in the end. The task of sentiment analysis can be carried out at different levels, i.e., at the level of the whole text, individual sentences, or aspect level. Sentiment analysis at the level of document assumes that reviews have an opinion about single entity [9, 10]. Analysis at sentence level determines tone for individual sentence or review (subjectivity analysis). Both document- and sentence-level analyses fail to find out the exact likings and disliking of reviewers.

Hence, aspect-level analysis is a more appropriate solution for a complete and accurate model. In this study, two datasets of approximately 800 reviews each are collected from Trip Advisor [1]; the reviews cover various hotels. These reviews are used as inputs to various machine learning algorithms (classifiers). In this study, we have focused on the class of linear classifiers. Classifiers are trained on features which are relevant to the classification tasks. These features are extracted from the data using various natural language processing (NLP) techniques. Finally, the performances of the trained classifiers are evaluated and compared [11].

2 Data Pre-processing

Data is seldom in the perfect form for direct use in machine learning techniques. The first need is to curate and clean the data. The absence of pre-processing can drastically affect the accuracy and efficiency of the overall process. Before applying any of the sentiment mining methods, it is a common practice to perform data pre-processing. Pre-processing of the collected reviews is performed by using the natural language toolkit (NLTK) platform available on Python [12, 13].

2.1 *Removing Stopwords*

Stopwords do not impart any meaning to the text and thus are superfluous. They are removed using the corpus of stopwords provided in the NLTK.

2.2 *Eliminating Special Characters*

Special characters like [] {} ()/' should be removed as they are not involved in semantics. Regular expression (RE) operation module in Python has been used for the job.

2.3 *Stemming*

Stemming deals with the reduction of a word to its native or root word. It helps in providing a common representation for different syntactic forms of the same word. The corpus 'WorldNet' from NLTK [12, 13] in Python is used, which lemmatizes the words using WorldNet's built-in Morphy function from nltk.stem.wordnet.

2.4 *Part of Speech (POS) Tagging*

Human speech is made up of different parts like verbs, adjectives, and nouns. Humans can understand these naturally, but a computer cannot. Therefore, each token is tagged as a POS which would allow the model to use this information to provide better results.

2.5 *N-Gram*

An n-gram is an adjacent series of ‘n’ tokens in a text. The elements may be base pairs, syllables, words, phonemes, or letters according to the application. These n-grams are typically collected from a speech corpus or text. If the items in the n-gram are words, then these n-grams are also known as shingles [14].

N-grams of texts are broadly used for providing contexts in opinion mining, text mining, and natural language processing (NLP) tasks. They are fundamentally a set of co-occurring words within a given window, and word-level n-grams have been used in this study.

3 Methodology

In this paper, linear model classifiers have been employed which are present in Natural Language Toolkit Package in Python. The overall workflow is presented in Fig. 1.

3.1 *Ensemble Classifiers*

Ensemble learners are a collection of classification algorithms which categorize input information points by making subjective choices of their computations at each step. Bayesian averaging is the main ensemble technique, but error correction and bagging/boosting are also incorporated in various algorithms [15]. This paper analyzes such methods and gives details explaining the reasons behind the better performance of ensembles as compared to simple classifiers. Some earlier work related to ensemble learners is reviewed; there is an indication that AdaBoost does not overfit quickly [16].

The ensemble techniques are used in the direction of combining the several base estimators with a known learning algorithm to develop the ability to generalize or robustness over a single estimator [16].

3.2 *Bagging*

Bagging creates different classifiers when there is an imbalance in the base learning algorithm, i.e., there is substantial change even when the input changes minimally. Bagging can be viewed as ways of exploiting the volatility in learning algorithms to improve classification accuracy [17].

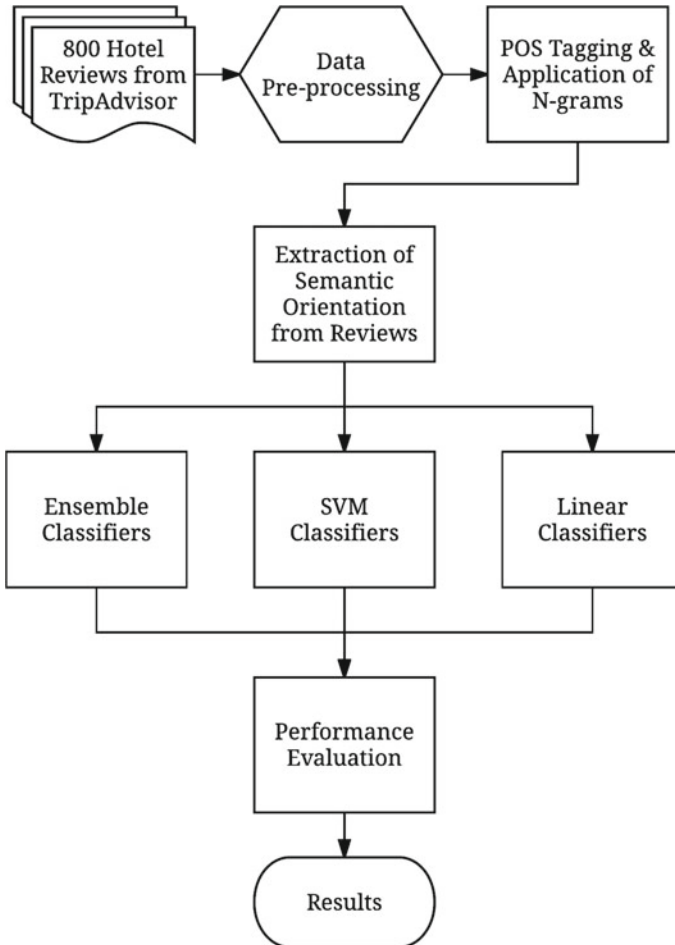


Fig. 1 Flow diagram of the proposed model

Bagging is also known as bootstrap aggregating. It helps in improving the accuracy of statistical classification and regression. The numerous editions are developed by considering bootstrap replicates and using each as new learning sets. It has been shown that bagging gives significant improvement in accuracy by performing tests on real and simulated datasets using regression trees [16].

3.3 Boosting

Boosting is an ensemble meta-algorithm intended for mainly controlling bias and for translating weak learners to strong ones. In boosting algorithms many weak learners are iteratively added and removed to get the final strong classifiers [16].

3.4 Naïve Bayes Classifier

The Bayesian classifier symbolizes a supervised statistical method for categorization. It is wholly based on Bayes theorem with independent assumptions between predictors. A naive Bayesian model is simple to construct, with simple iterative parameter evaluation making it mainly helpful for massive datasets. Despite its ease, the naive Bayesian classifier frequently performs well and often outperforms sophisticated classifiers [18].

Bayes theorem offers a method of evaluating the posterior probability, $P(c/p)$, from $P(c)$, $P(p)$, and $P(p/c)$. Naïve Bayes classifier assumes that the resultant value of a predictor (p) based on a class is not dependent on other predictors. The assumption is known as class conditional independence.

$$P(c/p) = \frac{P(p/c)P(c)}{P(p)} \quad (1)$$

Naïve Bayes classification is based on the Bayesian theorem equation. This can be simplified by an assumption that features are independent when the class values are provided. Despite this assumption, the classifier is successful in practice.

$$P(p/c) = \pi_{i-1}^n P(p_i/c) \quad (2)$$

where $p = \{p_1, \dots, p_i\}$ is a feature vector.

3.5 Support Vector Machines

SVMs are supervised models that contain a learning component which analyzes classification and regression analysis data. New inputs are then mapped into the same space and categorized it based on which side of the gap they fall in [19].

In our experiments, we have employed three classifiers from support vector machine; they are SVC, NuSVC, and linear SVC.

3.6 Linear Model Classifiers

Classification is the problem of recognizing which set of categories a new observation belongs to, in accordance with the training data containing labeled observations. In this paper, we have employed three types of linear model classifiers; they are SGD classifier, logistic regression classifier, and Bayes ridge classifier [6].

4 Performance Evaluation Measures

Evaluation metrics are the solution to understanding how your classification model performs when applied to a test dataset. When we use a classifier model for a problem, we almost always want to look at the correctness of that model as the number of correct predictions from all predictions made [20].

For the following measures, T_p represents true positives, F_p represents false positives, T_n is true negatives, and F_n denotes false negatives.

4.1 Accuracy

Accuracy is a measure used for evaluating the performance of categorization techniques. Accuracy values are less affected by variations in the number of correct decisions than precision and recall. The fraction of test set reviews is correctly classified:

$$A = \frac{T_p + T_n}{T_p + F_p + F_n + T_n} \quad (3)$$

4.2 Precision

This metric measures the correctness of a classifier. A lower precision means false positives, while higher precision means less false positives. Precision is defined as the number of true positives over the sum of a number of true positives and the number of false positives.

$$P = \frac{T_p}{T_p + F_p} \quad (4)$$

4.3 Recall

This metric measures the completeness, or understanding, of a classifier. Lower recall means a higher number of false negatives, while higher recall means less number of false negatives. Improving recall will reduce precision because it becomes progressively harder to be precise as the model space increases. The recall is defined as the number of true positives over the number of true positives plus the number of false negatives.

$$R = \frac{T_p}{T_p + F_n} \quad (5)$$

4.4 F-Measure

This metric was generated by the combination of two parameters that are precision and recall, and the single generated parameter is known as F-measure. F-measure is the weighted harmonic mean of recall and precision:

$$F1 = 2 \frac{P \times R}{P + R} \quad (6)$$

5 Experimental Results

The experiments are performed on a dataset of over 800 raw publicly available hotel reviews. The open-source implementation of pre-processing techniques and classifiers is used from NLTK library. The experiment is conducted using five ensemble learners, three support vector machine-based individual learners, and three naïve Bayes individual learners. All learners are used with unigram, unigram + bigrams, and unigram + bigrams + trigrams feature extraction methods. The comparative performance of each class of learner is plotted in graphs. Figure 2 compares the accuracy of the classifiers, whereas Fig. 3 compares the F-measure of the four different class of classifiers.

Accuracies of all the classifiers are presented in the form of graphs in Fig. 2. Random forest from the class of ensemble classifiers outperforms all other techniques as clearly seen from Fig. 2. Among the different classifiers, multinomial naïve Bayes classifier performance matches that of random forest. But due to the fact that multinomial naïve Bayes technique performs well when scaled to large datasets, it can be considered to perform best among the discussed classifiers.

F-measure is a score which gives us an understanding of the actual performance in terms of precision and recall of the classification. As seen from Fig. 3, again multinomial naïve Bayes and random forest techniques produce the best results.

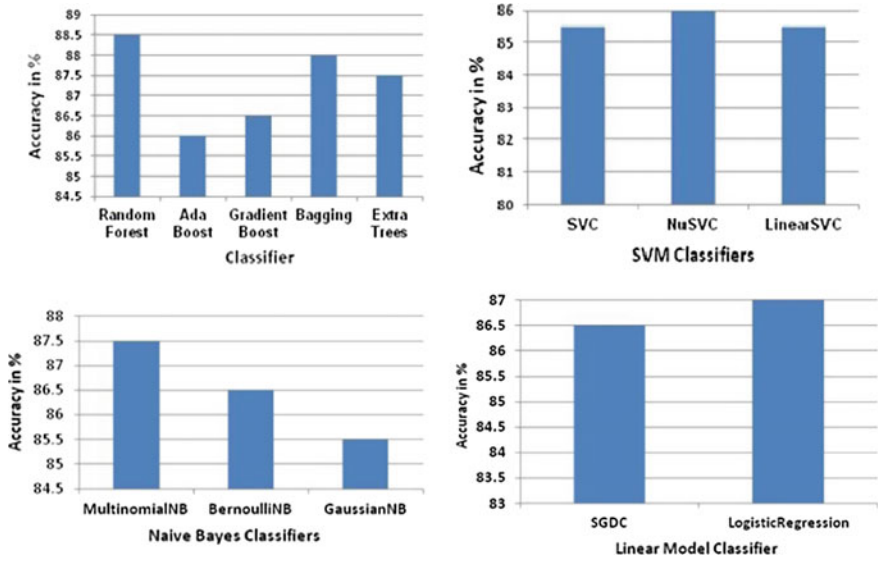


Fig. 2 Accuracy comparison of the classifiers

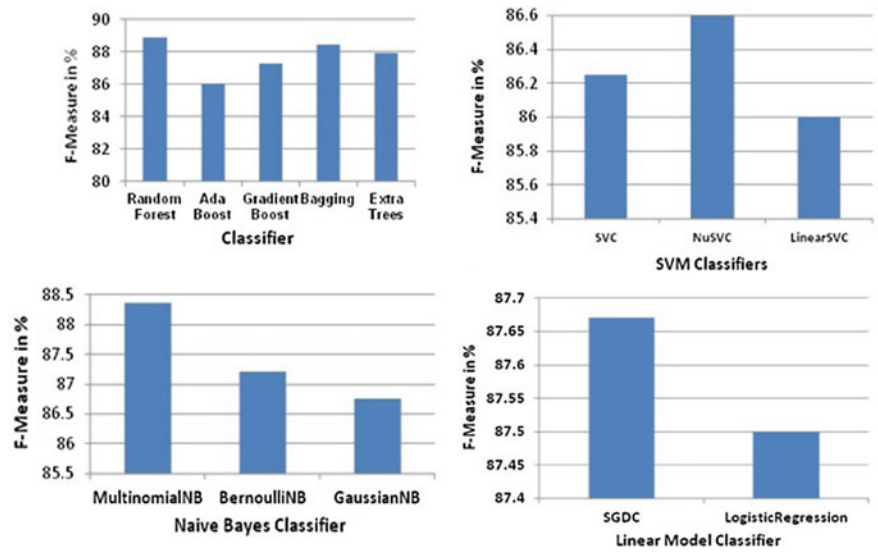


Fig. 3 Comparison of F-measure of classifiers

Other classifiers also have good F-measure scores, but multinomial naïve Bayes technique edges ahead of the rest.

6 Conclusion and Future Work

Above results compare ensemble methods, SVM, linear models, and naïve Bayes classifier. In ensemble methods, bagging classifier is the most effective classifier using unigram and bigram classifications, whereas random forest classifier is the most effective classifier while using uni-, bi-, and trigrams. In support vector machine methods, SVC is the best when using unigram classification.

In naïve Bayes classifiers, multinomial classifier is the most efficient classifier using unigram, bigrams, and trigrams. In linear models, stochastic gradient descent classifier (SGDC) is the best when using unigram, bigrams, and trigrams.

Overall multinomial naïve Bayes classifier is the most efficient classifier, and this is consistent with our knowledge that naïve Bayes classifiers are most productive on the small datasets. To implement this project, data was collected manually, so it has just 800 reviews. Naïve Bayes classifier is the most efficient classifier for this dataset, although the remaining classifiers also give fair results.

Comparisons of standard classifiers have been presented in this study, but this is not an exhaustive comparison. In recent years, there has been a rise of neural network classification, especially using deep networks. This work can be extended toward deep learning which has the potential of outperforming all the classification techniques discussed here. The main hurdle to the application of deep neural networks for classification tasks is lack of data. But with the proliferation of data in past few years, next step can be utilizing huge datasets to extend this work to include deep learning.

References

1. Trip Advisor. <https://www.tripadvisor.in/>.
2. O'Connor, P. (2008). User-generated content and travel: A case study on Tripadvisor.Com. In *Information and communication technologies in tourism 2008* (pp. 47–58).
3. Sisodia, D. S., & Reddy, N. R. (2017). Sentiment analysis of prospective buyers of mega online sale using tweets. In *2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPSI)* (pp. 2734–2739). IEEE, 2017.
4. Sisodia, D. S., & Reddy, R. (2019). Analysis of public sentiments about mega online sale using tweets on big billions day sale. In *Sentiment Analysis and Knowledge Discovery in Contemporary Business* (pp. 59–76). IGI Global, 2019.
5. Bjørkelund, E., Burnett, T. H., & Nørvåg, K. (2012). A study of opinion mining and visualization of hotel reviews. In: *Proceedings of the 14th International Conference on Information Integration and Web-Based Applications & Services—IIWAS '12*, ACM, p. 229.
6. Yuan, G. X., Ho, C. H., & Lin, C. J. (2012). Recent advances of large-scale linear classification. *Proceedings of the IEEE*, 100, 2584–2603.

7. Kasper, W., & Vela, M. (2011). Sentiment analysis for hotel reviews. In: *Proceedings of the Computational Linguistics-Applications Conference* (pp. 45–52).
8. Shi, H. X., & Li, X. J. (2011). A sentiment analysis model for hotel reviews based on supervised learning. *Proceedings—International Conference on Machine Learning and Cybernetics*, 3, 950–954.
9. Pan, W., Shen, X., & Liu, B. (2013). Cluster analysis: Unsupervised learning via supervised learning with a non-convex penalty. *Journal of Machine Learning Research: JMLR*, 14, 1865.
10. Bross, J. (2013). *Aspect-oriented sentiment analysis of customer reviews using distant supervision techniques*. PhD dissertation.
11. Gräbnera, D., & Zankerb, M. (2012). Classification of customer reviews based on sentiment analysis. In: *19th Conference on Information and Communication Technologies in Tourism* (pp. 1–12). Berlin: Springer.
12. Bird, S., & Loper, E. (2004). NLTK: The natural language toolkit. In: *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics* (pp. 1–4).
13. Natural Language Toolkit. <http://www.nltk.org/>.
14. Jurafsky, D., & Martin, J. H. (2016). Language modeling with N-grams. In: *Speech and language processing* (pp. 1–28). London: Pearson Education.
15. Sisodia, D. S., & Yogi, A. K. (2019). Performance evaluation of ensemble learners on smartphone sensor generated human activity data set. In *Data, engineering and applications* (pp. 277–284). Singapore: Springer.
16. Gopika, D., & Azhagusundari, B. (2014). An analysis on ensemble methods in classification tasks. *International Journal of Advanced Research in Computer and Communication Engineering*, 3, 7423–7427.
17. Bryll, R., Gutierrez-Osuna, R., & Quek, F. (2003). Attribute bagging: Improving accuracy of classifier ensembles by using random feature subsets. *Pattern Recognition*, 36, 1291–1302.
18. Narayanan, V., Arora, I., & Bhatia, A. (2013). Fast and accurate sentiment classification using an enhanced Naive Bayes model. In *International Data Engineering and Automated Learning, Lecture Notes in Computer Science* (Vol. 8206, pp. 194–201).
19. Zainuddin, N., & Selamat, A. (2014). Sentiment analysis using support vector machine. In: *2014 International Conference on Computer, Communications, and Control Technology (I4CT)* (pp. 333–337).
20. Huang, G., Song, S., Gupta, J. N. D., & Wu, C. (2014). Semi-supervised and unsupervised extreme learning machines. *IEEE Transactions on Cybernetics*, 44, 2405–2417.
21. Dietterich, T. G. (2000). Ensemble methods in machine learning. In *International workshop on multiple classifier systems* (pp. 1–15). Berlin, Heidelberg: Springer.

Proposed Data Structure for Storage of Metrics Values: Misuse Case Oriented Quality Requirements (MCOQR) Framework Perspective



Sunita Choudhary, C. Banerjee, Ajeet Singh Poonia, Arpita Banerjee and S. K. Sharma

Abstract Security may be quantified to measure the level of software security implementation. These quantified measures are called security metrics. These metrics need to be stored in some central or local repository for further analysis and refinement of security of software proposed to be developed or modified. Through the research work carried out by the researchers, it is proposed to have a data structure for storage of metrics values which are identified and collected using the Misuse Case Oriented Quality Requirements (MCOQR) framework. The research work highlights and discusses the internal arrangement of various data sets in the data structure and their relationship with each other. The data sets thus proposed are properly synchronized with the industry accepted standards like Common Vulnerability Scoring System (CVSS), Common Vulnerability Enumeration (CVE), and Common Weakness Enumeration (CWE). The work proposed is an extension of Misuse Case Oriented Quality Requirements (MCOQR) framework and metrics and includes software application-specific database. The research work also highlights the areas where further research work can be carried out to further strengthen the entire system.

Keywords Security data sets · Misuse cases · CVSS · CVE · MCOQR metrics

S. Choudhary · A. S. Poonia
Government College of Engineering and Technology, Bikaner, India
e-mail: sunitadangi@gmail.com

A. S. Poonia
e-mail: pooniaji@gmail.com

C. Banerjee (✉)
Amity University Rajasthan, Jaipur, India
e-mail: chitreshh@yahoo.com

A. Banerjee
St. Xavier's College, Jaipur, India
e-mail: arpitaa.banerji@gmail.com

S. K. Sharma
Modern Institute of Technology and Research Center, Alwar, India
e-mail: sharmasatyendra_03@rediffmail.com

1 Introduction

The dependence of the human world on the use of software has increased by the invention of new technologies. In the present time, the software has become an essential and integral part of day-to-day life [1, 2]. The economy on the global level is reliant on the coordinated and secure implementation and use of the software. It has penetrated the human life so much so that the trust level of people is influenced and judged by the security level of software [3, 4].

The engineering team will use the proposed framework and associated algorithms to identify, define, and record cases of abuse in the repository according to the proposed work. A central repository of CVSS, CVE, Misuse Cases and Application-Specific Misuse Cases is available for this purpose [5, 6].

There is a prerequisite that the data available in the central repository should be finely tuned and updated on a regular basis with CVSS and CVS external repositories available online. As and when any misuse cases will be identified, it shall be recorded in misuse cases as well as application-specific misuse cases database [7].

On the basis of the data available in the central repository, a proposed algorithm will be used to retrieve, calculate, and fill in the worksheet misuse case modeling. The proposed metrics for Misuse Case Oriented Quality Requirements (MCOQR) will apply to these populated data to derive the final data containing predicted misuse case counts, scoring and ranking disclosing interrelated multidimensional levels of threat source, threat impact, level of countermeasures, and dominant vulnerability type.

The final outcome will be in form of a worksheet showing various levels of indicators and estimators in varied colors (for identification purpose) which may be used by the security requirements engineering team to analyze and interpret the level of security implementation in the software application well before its design and development. This shall also enable the team to further strengthen the security aspect of software by removing the defects of misuse case modeling with the introduction of more countermeasures.

2 Proposed Data Structure for MCOQR Metrics

The proposed algorithm for identification of vulnerabilities and associated misuse case using CVSS, CVE standards figures out the following data structures:

- a. A central repository consisting of Vulnerability db, CVE db, Misuse Case db, and CVSS Metrics db which needs to be updated from time to time using CVE repository and CVSS latest version repository.
- b. A repository consisting of Application-Specific Misuse Case db (Fig. 1).

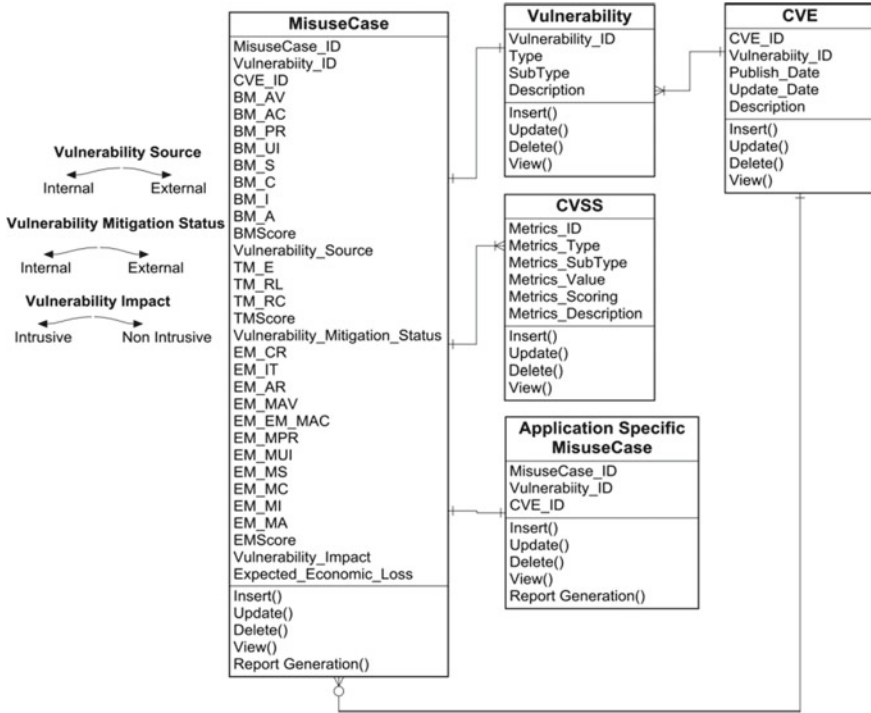


Fig. 1 Showing the various data tables of MCOQR metrics framework along with their relationships

3 Implementation Mechanism

Once the business objectives are identified and properly aligned with the asset identified, the next step is to identify the associated security risks and objectives. The various loopholes are studied and identified using the identified security risk and objective and become the step for vulnerability identification. This step is very important for further processing and requires expertise because the entire process will become futile if the correct identification of vulnerability is not carried out.

There is a central repository consisting of the CVE Database for Common Vulnerability Enumeration, the Vulnerability Database, the Common Vulnerability Document, the Misuse Case Database and a repository containing only the Application-Specific Misuse Case Database. The security team is responsible for updating the CVE common vulnerability listing database from time to time with the external CVE repository and the CVSS document with the external CVSS repository.

Once the vulnerability has been identified, the CVE database will be used to categorize it. We have taken two main vulnerability categories for our research purposes, i.e., database vulnerability and application vulnerability. The framework has been designed to accommodate more vulnerability categories. Once the vulnerability



is classified as vulnerability in the database or vulnerability in the application, the associated record is written/updated in the database of vulnerability. In this way, all possible vulnerabilities should be identified and stored in the vulnerability database for this software application.

After the subclassification has been completed, the associated vulnerability is added/updated to the misuse case database residing in the central repository, and a copy of the application-specific misuse case database is attached. These data can also be used to create worksheets for CVSS and MCOQR, and the security team can use the worksheet created to analyze defects in the modeling of misuse cases using the six proposed security indicators and estimators. This modeling can be used to specify security requirements during the engineering phase of requirements and can also help to further improve the safety of the system to be constructed.

4 Results and Validation

This proposed data structure was applied to an industry real-life project (the identity is hidden at the company's request), and the final result of the security assessment is calculated in accordance with the prescribed implementation mechanism.

The level of security assurance is then compared to the security assurance of the other project, which did not apply the proposed algorithm. The study shows that the risk level is reduced by up to 42.5%. We do not provide details of the validation results in this paper because of the page limit restriction; we will discuss them in our next paper.

5 Conclusion and Future Work

We proposed a structured way to organize identified and classified vulnerabilities and associated cases of misuse in a central repository and a specific application repository. We also promote awareness of security among the various stakeholders involved in the development process through our work. Future work may include the validation of the proposed Misuse Case Oriented Quality Requirements framework (MCOQR) and derived metrics with a large set of data.

References

1. Taylor, P. (2015). The importance of information and communication technologies (ICTs): An integration of the extant literature on ICT adoption in small and medium enterprises. *International Journal of Economics, Commerce and Management*, 3(5).

2. Banerjee, C., & Pandey, S. K. (2009). Software security rules. *SDLC Perspective*. arXiv preprint: 0911.0494.
3. Chess, B., Do, A., Fay, S., & Thornton, R. (2016). *U.S. Patent No. 9,400,889*. Washington, DC: U.S. Patent and Trademark Office.
4. Roztocki, N., & Weistroffer, H. R. (2015). Information and communication technology in transition economies: An assessment of research trends. *Information Technology for Development*, 21(3), 330–364.
5. Poonia, A. S., Banerjee, C., Banerjee, A., & Sharma, S. K. (2018). Vulnerability identification and misuse case classification framework. In *Soft computing: Theories and applications* (pp. 659–666). Singapore: Springer.
6. Banerjee, C., Banerjee, A., & Sharma, S. K. (2018). Estimating influence of threat using Misuse Case Oriented Quality Requirements (MCOQR) metrics: Security requirements engineering perspective. *International Journal of Hybrid Intelligence Systems (IJHIS)*, 14(1–2), 1–11. USA: IOS Press, MIRS Lab. ISSN 1875-8819.
7. Banerjee, C., Banerjee, A., Poonia, A. S., Sharma, S. K., & Pandey, S. K. (2018). Improving the results of Misuse Case Oriented Quality Requirements (MCOQR) framework metrics: Secondary objective perspective. In *Proceeding of International Conference on Soft Computing: Theories and Applications SoCTA 2017, Advances in Intelligent Systems and Computing*, New York: Springer Publication.

Comparative Analysis of Hindi Text Summarization for Multiple Documents by Padding of Ancillary Features



Archana N. Gulati and Sudhir D. Sawarkar

Abstract There is an enormous amount of textual material, and it is only growing every single day. The data available on Internet comprised of Web pages, news articles, status updates, blogs which are unstructured. There is a great need to reduce much of these text data to shorter, focused summaries that capture the salient details so that the user can navigate it more effectively as well as check whether the larger documents contain the information that we are looking for. Text summary is generating a shorter version of the original text. The need of summarization arises because every time it is not possible to read the detailed document due to lack of time. Automatic text summarization methods are greatly needed to address the ever-growing amount of text data available online both to better help discover relevant information and to consume relevant information faster. To address the issue of time constraint, an extractive text summarization technique has been proposed in this research work which selects important sentences from a text document to get a gist of information contained in it. A fuzzy technique has been used to generate extractive summary from multiple documents by using eight and eleven feature sets. The eleven feature set combines the existing eight features (term frequency-inverse sentence, length of sentence in the document, location of sentence in document, similarity between sentences, numerical data, title overlap, subject object verb (SOV) qualifier, lexical similarity) and three ancillary features (proper nouns, hindi cue phrase, thematic words). It was seen that applying fuzzy technique with eleven features gave better results for summarization than the same using eight features. The precision increases in the range of 3–5% for different datasets. Datasets used were Hindi news articles from online sources.

A. N. Gulati (✉)

Department of Computer Engineering, Datta Meghe College of Engineering, Mumbai, Maharashtra, India

e-mail: ang.cm.dmce@gmail.com

S. D. Sawarkar

Datta Meghe College of Engineering, Mumbai, Maharashtra, India

e-mail: sudhir_sawarkar@gmail.com

Keywords Multiple documents · Ancillary features · Extraction · Fuzzy technique · Ancillary features · Hindi news articles

1 Introduction

There are varieties of online resources for Hindi news articles. Every latest news has a reference in multiple newspapers. Moreover, certain news span over multiple days in various newspapers. So, if anybody wants a gist of all this news at one place, there should be some method to achieve this. Text summarization systems come into picture at this point.

Such systems generate a reduced version of the original text which makes the task simpler and easier for reader with lack of time. There are two ways in which summarization can be achieved either extractive (picking up important sentences) or abstractive (creating more generalized text). Further, summarization can be applied to documents written in different languages like English, Hindi, Marathi, Kannada, etc., or for that matter any regional language, although implementing it in one language might be easier than the other. Also different algorithms have been devised to generate such summaries.

In the proposed system, an extractive text summarization method using fuzzy logic is defined which works for Hindi text documents. Two sets of features are taken for summarization purpose—one set of eight basic features (term frequency-inverse sentence, length of sentence in the document, location of sentence in document, similarity between sentences, numerical data, title overlap, subject object verb (SOV) qualifier, lexical similarity [1]) and a second set of eleven features. The set of eleven features consists of the eight basic features from eight feature set and three ancillary features (proper nouns, Hindi cue phrase, thematic words) [2]. The purpose of this work was to compare the two feature sets and check the improvement in precision after adding the three ancillary features to the existing eight feature set.

Hindi being the national language of India was selected for the study purpose. Majority of nationals read Hindi news; hence, it was decided to work on the national language and help them to get the gist of Hindi news in one go. The summarization technique makes use of a mix of statistical and semantic features, statistical features such as term frequency-inverse sentence frequency, sentence length, and sentence position, whereas semantic features like SOV qualification, sentence similarity and likewise.

2 Related Work

Yogesh Kumar Meena and Dinesh Gopalani in their paper have proposed an extractive text summarization technique using various evolutionary algorithms. They have defined a wide variety of features to be used by the fitness function for summary generation [3].

Udo Hahn and Inderjeet Mani have come up with the challenges of automatic summarization. They have considered both the methods of summarization, i.e., Extractive and abstractive, and discussed the challenges like the reduction rate and evaluation criteria for the summaries generated. According to the authors, the methods of creating and evaluating a summary must complement each other. Secondly, they say a lot of background knowledge is needed to achieve high reduction. Most of the methods apply linear weighting models which weight individual sentences for different features and then find the overall weights by summing up the individual weights. These weights will help in deciding which sentences to include in the summary [4].

S. Santhana Megala, Dr. A. Kavitha, and Dr. A. Marimuthu in their paper have proposed a single document extractive text summarization technique using fuzzy logic and neural network. They have compared both the methods by experimenting on 50 different legal documents and have concluded that fuzzy logic gives a better precision than neural network [5].

F. Kyoomarsi, H. Khosravi, E.Eslami, and M. Davoudi have created a fuzzy logic analyzer that calculates relevance score for each sentence, and those sentences having relevance score above the predefined threshold are selected to be included in summary. The method is a combination of fuzzy logic and WordNet. The process continues till a definite compression ratio is achieved [6].

Yogesh Kumar and Dinesh Gopalani have proposed a feature priority-based filtering method where sentence location is given highest priority while filtering out redundant sentences, and later on features like TF/ISF, named entity, and proper nouns are used one by one rank wise. According to them, sentence position is a very good parameter for extracting sentences [7].

S. A. Babar and Pallavi D. Patil have devised a method for single document extractive text summarization which uses fuzzy logic system. But here the difference is that latent semantic analysis (LSA) to extract semantic relations between concepts from the text documents has been used. So adding LSA to existing fuzzy logic system improves the quality of the summary [2].

Chetana Thaokar and Dr. Latesh Malik in their paper describe a technique to generate extractive summary for single Hindi document using statistical and linguistic approach to find most relevant sentences. This summarization system uses total eight features for calculating sentence score and uses genetic algorithm for ranking and optimization of the sentences [1].

Pallavi D. Patil and P. M. Mane in their work have combined latent semantic analysis and fuzzy logic together to improve performance of single as well as multi-document text summarization. LSA identifies semantically comparable words and sentences. Fuzzy logic is used to remove uncertainties in the given problem [8].

Patil Pallavi and Mane P. M. in their paper have described the various text categorization approaches and have also focused on how fuzzy logic and latent semantic analysis techniques can be used to build summarization system [9].

Pallavi Patil and N. J. Kulkarni define a text summarization approach with stages such as pre-processing, feature extraction, fuzzy logic scoring, and sentence selection. The important features used are sentence position, sentence length, title word, numerical data, thematic words, and sentence-to-sentence similarity [10].

S. Santana Megala and A. Kavitha have proposed a method basically for preparing headnotes automatically from legal documents. Generating headnotes from legal documents is a very tedious and time-consuming job for judges and advocates for making important decisions and judgments. So in the proposed work, a headnote generation technique using fuzzy logic has been defined [11].

3 Proposed Methodology

In this section, the text summarization using eight and eleven feature sets is described and then a comparison using these two different feature sets is carried out. The basic flow of the system is shown in Fig. 1.

3.1 Pre-processing

Pre-processing is carried out to prepare the text documents for analysis and includes steps such as segmentation, tokenization, stop words removal, and stemming.

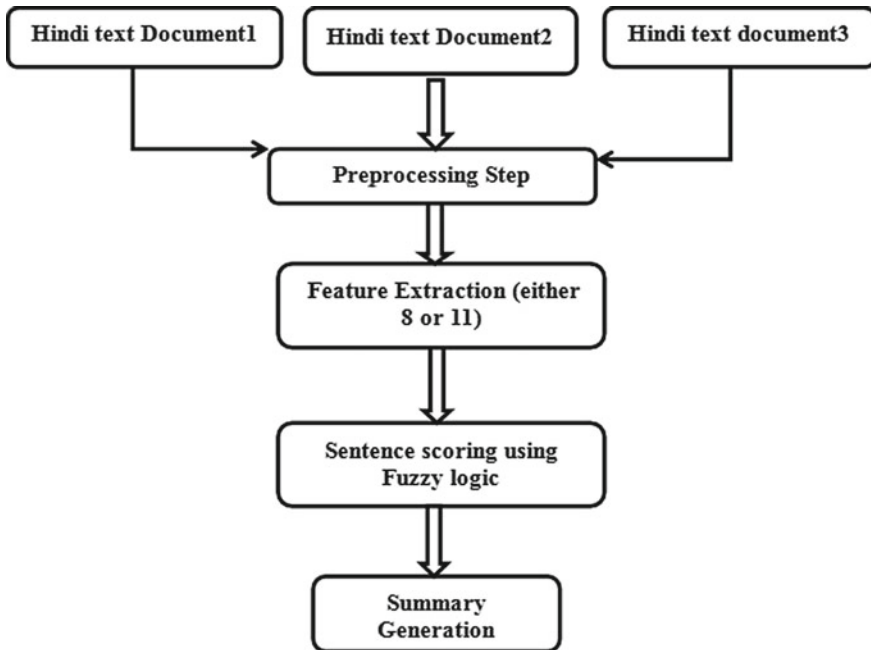


Fig. 1 Basic flow of the system

3.2 Feature Extraction

The eleven feature set has the original eight features [1] described in eight feature set and three ancillary features [2]. The decision of whether to include a sentence in summary or not depends upon the score of these features for individual sentences. The basic eight features [1] used are:

Term frequency-inverse sentence frequency (TeFe/InvSeFe): This feature is very important to eradicate the impact of frequently occurring terms which are not actually so useful in the final summary.

Length of sentence in the document (SeLe): This feature helps us in extracting sentences having length between 5 and 20. Sentences with words less than 5 do not carry significant information, and those greater than 20 tend to contain redundant information; hence, they need not be considered in summary.

Location of sentence in document (LS): This feature is used to identify important sentences to be included in summary depending upon a threshold value. Normally, few sentences in the beginning and end of a document tend to carry important information.

Similarity between sentences (Sim(Si, Sj)): Sentence similarity is computed as number of words in a sentence matching with all other sentences in the document. More the similarity, more the weightage.

Numerical data (Nd): Sentences containing some numbers like important dates, quantity, financial information, etc., may convey object verb (SOV) qualifier: For calculating this feature, POS tagging for each word in the sentence is done using the CoreNLP tool developed by Stanford. The first noun word in the sentence is marked as subject and the subsequent noun as object of the sentence. Further, the sentence is parsed to search for a verb, and if found the sentence is said to be SOV qualified and given a higher weight. Other sentences which do not qualify are given a lower weight.

Lexical similarity: As against sentence similarity which only calculates the word-to-word similarity between sentences, this feature tries to find out the semantic similarity by trying to find out the synonyms, hypernyms, and hyponyms by using World-Net support.

The three ancillary features [2] added in the eleven feature set are:

Proper nouns: Existence of a proper noun indicates that the sentence may contain some important information about the subject. Hence, more the number of proper nouns occurring in the sentence, the higher will be the weightage.

Hindi Cue Phrase: Certain words are more emphasizing in the sentences or they highlight some important point in the sentence. Such words are cue phrases, and their existence adds more meaning to the sentence. Certain Hindi cue phrases are उद्घाटन, घोषणा, आधारपर, महत्वपूर्ण, एलान etc.

Thematic words: For this feature, frequently occurring words in the document are identified and a threshold of n words is decided. The sentences having maximum of these thematic words are considered as more important.

3.3 Summarization Using Fuzzy Logic

Fuzzy logic can handle both numeric and linguistic data. It is used to imitate the process of human reasoning. The same concept has been used here for text summarization. First, the eleven features were extracted. Then the fuzzer is fed separately with eight features first and then with eleven features. The membership function used is the triangular membership function.

The input membership function for each feature is divided into five fuzzy sets which are composed of unimportant values, i.e., low (L) and very low (VL), medium (M), and important values, i.e., high (H) and very high (VH). Along with the features extracted, the if-then-else rules defined in the knowledge base are used to decide the degree of importance of the sentence on the basis of which they are ranked and thus a decision is made whether to include the sentence in the final summary or not [12].

4 Implementation Details

At most three Hindi news articles on the same topic are taken at a time. Then the pre-processing steps are applied on those documents. After pre-processing, the eleven feature values are calculated from F1 to F11. The final score is calculated for each Hindi document separately for eight feature set and eleven feature set. After getting final score, those sentences having highest score are selected to be included in the final summary. After getting summary of each Hindi document, all the summaries are combined and again feature extraction step is applied on the final summary separately for both the feature sets.

The sentences with highest score from this metadocument are selected for inclusion in final summary. So finally, two summaries are generated for the same three input Hindi text documents, i.e., one for eight feature set and one for eleven feature set. Sample dataset used has been shown below.

5 Results and Analysis

Finally, a comparative analysis is carried out on both the summaries generated with the two different feature sets. The evaluation criteria used are precision, recall, and F-score.

5.1 Precision

Precision is number of correct sentences divided by number of sentences extracted.

5.2 Recall

Recall is number of correct sentences divided by the number of sentences that should have been extracted.

5.3 F-Score

F-score is the harmonic mean of precision and recall.

The summaries generated by both the methods are compared with expert human generated summaries. It is found that summary generated using eleven feature set gives a better summarization as compared to eight feature set. This implies that adding the three ancillary features, i.e., proper nouns, cue phrases, and thematic features, improves the summarization process.

Figure 2 shows the values of precision, recall, and F-score using eight features, and Fig. 3 shows values for the same with eleven features for a sample dataset.

Table 1 shows the comparative analysis of precision, recall, and F-score for eight and eleven feature set over multiple datasets and then a final average value for comparison purpose.

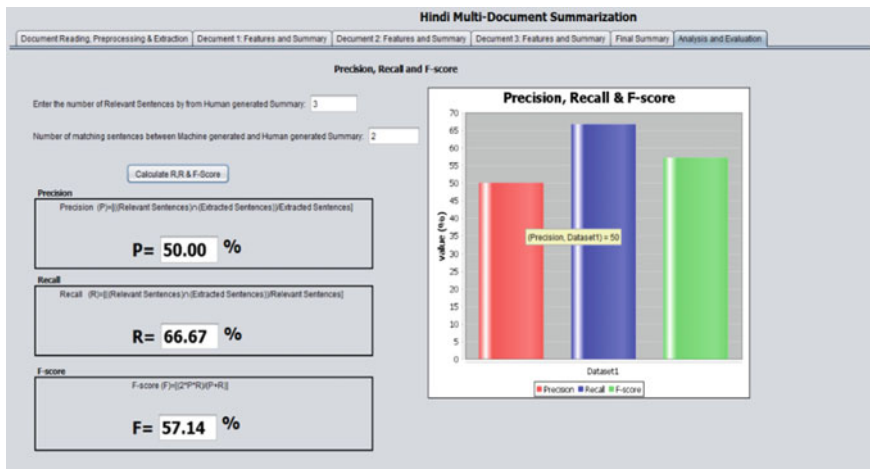


Fig. 2 Precision, recall, and F-score using eight feature set

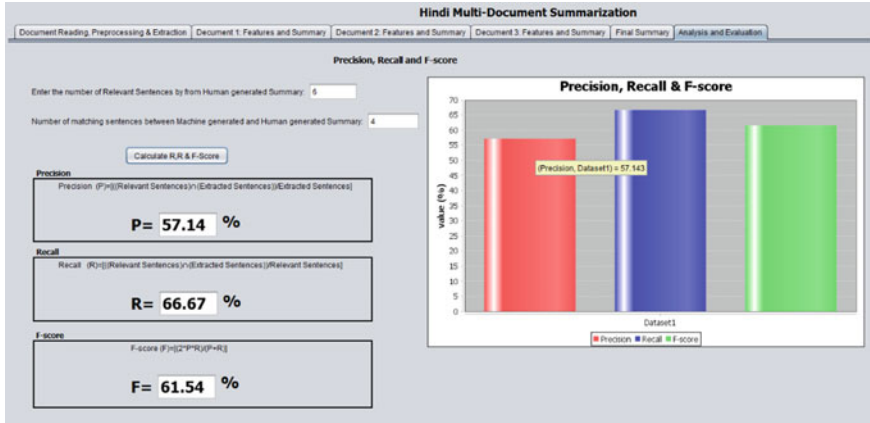


Fig. 3 Precision, recall, and F-score using eleven feature set

Table 1 Comparison of eight and eleven feature set values of precision, recall, and F-score for five different datasets

| | Eight feature set | | | Eleven feature set | | |
|-----------|-------------------|--------|---------|--------------------|--------|---------|
| | Precision | Recall | F-score | Precision | Recall | F-score |
| Dataset 1 | 50.00 | 66.67 | 57.14 | 57.14 | 66.67 | 61.54 |
| Dataset 2 | 66.67 | 56.66 | 61.25 | 69.24 | 60.66 | 64.67 |
| Dataset 3 | 75.00 | 69.66 | 72.23 | 74.54 | 68.96 | 71.64 |
| Dataset 4 | 66.67 | 50.00 | 57.14 | 64.00 | 56.12 | 59.80 |
| Dataset 5 | 71.43 | 62.50 | 66.67 | 78.26 | 66.56 | 71.94 |
| Average | 65.95 | 61.10 | 62.89 | 68.64 | 63.80 | 65.92 |

6 Findings and Conclusion

Thus, a fuzzy technique was applied over multiple documents to generate extractive summary by using eight and eleven feature sets. On comparison, it was seen that applying fuzzy technique with eleven features (base eight features and three ancillary features) gave better results than the same using eight features. The increase in precision varies in the range of 3–5% for different datasets. Datasets used were Hindi news articles from online sources.

7 Future Work

Currently, one fuzzy-based algorithm has been implemented on Hindi text for both eight feature set and eleven feature set, proving eleven feature set to give better



precision. Further experimentation would be carried out on other algorithms like neural networks for the two feature sets. Also work is planned for multiple languages.

References

1. Thaokar, C., & Malik, L. (2013). Test model for summarizing Hindi text using extraction method. In *IEEE Conference on ICT 2013*.
2. Babar, S. A., & Patil, P. D. (2015). Improving performance of text summarization. In *International Conference on Information and Communication Technologies (ICICT 2014)*, *Procedia Computer Science* (Vol. 46, pp. 354–363).
3. Meena, Y. K., & Gopalani, D. (2015). Evolutionary algorithms for extractive automatic text summarization. In *International Conference on Intelligent Computing, Communication & Convergence (ICCC-2014)*, *Procedia Computer Science* (Vol. 48, pp. 244–249).
4. Hahn, U., & Mani, I. (2000). The challenges of automatic summarization. In *2000 IEEE*.
5. Megala, S. S., Kavitha, A., & Marimuthu, A. (2014). Enriching text summarization using fuzzy logic. (*IJCSIT*) *International Journal of Computer Science and Information Technologies*, 5(1), 863–867.
6. Kyoomarsi, F., Khosravi, H., Eslami, E., & Davoudi, M. (2010). Extraction based text summarization using fuzzy analysis. *Iranian Journal of Fuzzy Systems*, 7(3), 15–32.
7. Kumar, Y., & Gopalani, D. (2015). Feature priority based sentence filtering method for extractive automatic text summarization. In *ICCC-2015*, *Procedia Computer Science* (Vol. 48, pp. 728–734).
8. Patil, P. D., & Mane, P. M. (2015). Improving the performance for single and multi-document text summarization via LSA & FL. *IJCST*, 2(4).
9. Patil, P. D., & Mane, P. M. (2014). A comprehensive review on fuzzy logic & latent semantic analysis techniques for improving the performance of text summarization. *International Journal of Advance Research in Computer Science and Management Studies (IJARCSMS)*, 2(11).
10. Patil, P. D., & Kulkarni, N. J. (2014). Text summarization using fuzzy logic. *International Journal of Innovative Research in Advanced Engineering (IJIRAE)*, 1(3).
11. Santana Megala, S., & Kavitha, A. (2014). Feature extraction based legal document summarization. *IJARMS*, 2(12).
12. Suanmali, L., Salim, N., & Binwahlan, M. S. (2009). Fuzzy logic based method for improving text summarization. *International Journal of Computer Science and Information Security (IJCSIS)*, 2(1).

RC-Network and Comm-Network for Improvement of Research Collaboration and Communication Among Delhi University Teachers



Narender Kumar, Sapna Malhotra, Chitra Rajora and Ravins Dohare

Abstract Research collaboration studies have become significant indicators of research, productivity, and developments of organizations. In the research work, the authors have proposed and construct research collaboration network (RC-network) and communication network (Comm-network) for the Delhi University teachers. Teachers are treated as vertices, and two vertices are connected if they have written a research document together in the case of RC-network, whereas if they communicate (talk about education, politics, rights, or other socially relevant issues) with each other, then it is a case of Comm-network. In the research work carried out, the researchers have collected the data through different resources and questionnaires. There are 712 teachers' records of research document and 214 teacher's records from questionnaire in the collected data. Two networks are constructed as RC-network and Comm-network. The researchers have found through the study that both networks follow scale-free degree distribution as reported by the other similar study. These networks unpacked several hidden characteristics of collaboration and communication among teachers of Delhi University. Constructed networks are sparse and partitioned in huge number of connected components. Moreover, they have identified the key clusters of research collaborations and communications among teachers. These clusters consist of teachers mostly from the same college or department or center. This indicates that the research collaborations are not of interdisciplinary nature in Delhi University. Even the communications among the teachers from different disciplines or colleges take place less frequently than those among the teachers from the same

N. Kumar · S. Malhotra

Department of Mathematics, Gargi College, University of Delhi, New Delhi, India
e-mail: nkumariitd@gmail.com

S. Malhotra

e-mail: kaursapna@gmail.com

C. Rajora

Department of Commerce, Gargi College, University of Delhi, New Delhi, India
e-mail: chitra.rajora@gmail.com

R. Dohare (✉)

Center for Interdisciplinary Research in Basic Sciences, Jamia Millia Islamia, New Delhi, India
e-mail: ravinsdohare@gmail.com

discipline or college. This study may prove to be helpful in improving and understanding the research and communication among teachers of Delhi University.

Keywords RC-network · Comm-network · Path lengths · Clustering coefficient · Clusters

1 Introduction

Science and technology consist of a broad range of activities like fundamental research, scholar activity, basic science, and development activities mainly focusing on the production of new products and process [1]. It has played a major role in the economic growth and development over the last 25 years; therefore, it can be treated as unseparated part of many countries and their innovation system [1, 2]. Innovation and competition are two key points for advancement of science and technology. Competition has its disadvantages but through research collaboration, these disadvantages can be transformed into advantages [3]. Research collaboration is defined as the interaction between two or more individuals in order to complete a particular task with common shared goal [4].

Collaboration among researchers usually can be seen in informal ways [5, 6]; therefore, many scientists made a clear difference between researcher's collaboration and communication [7–9]. They categorized manuscript coauthorship, presentation collaboration in conferences, seminars, workshops, and meetings under the formal category, whereas conversation, feedback from colleagues, journal editor, etc., under the informal category. Several collaboration networks were constructed and analyzed [10–13] as a coauthorship network. Some organization has been set research collaboration network as their aim. Collaboration networks are helpful in understanding the topological properties of large networks [13–15].

In recent time with the availability of large database and increase in mathematical and computational power, scientists used collaboration network to learn the process-generating network topology [16]. Network/Graph theory has been emerged as the most efficient tool that can be applied to construct, analyze, and understand the biological and sociological complex systems. According to the literature, it can be observed that coauthorship is not a sufficient indicator to measure the scientific collaboration; Bozeman and Corley [5] used self-reported data to prepare the collaboration network. For this work, they asked the participants to give the details of people with whom they were engaged in research since last one year. They discussed that some problems associated with publish-based work were solved such as listing the name of authors based on only social relation [6, 17]. They also discussed disadvantages of self-reported data, for instance, accuracy of the collected data.

2 Methodology

The University of Delhi has different types of nature of appointment for teachers as permanent, temporary, ad hoc, and guest. In this article, we include all four categories of teachers, but research associates and students are not considered. We created two networks, namely RC-network and Comm-network. For both types of networks, we used different methods to collect the data. We used secondary data available on the Internet for the construction of RC-network.

We collected the teacher's data from official Web sites of the colleges of departments/centers of the university. Construction of RC-network is based on the data of 6252 teachers of 27 departments/centers and 58 colleges of the university. For Comm-network, we used the questionnaire in hard copy format. The questionnaire was two-page long having simple questions. The first page contains the information about respondent and on second page information of other teacher's was asked with whom respondent communicates. Initially, questionnaire was designed as an electronic document, but several teachers commented and suggested that paper-pen (hard copy) format is much easier and comfortable to fill.

Therefore paper-pen format was used to collect the data through questionnaire. A team of students worked and collected the data from different colleges or departments/centers of the university. The teachers were requested to fill questionnaire without any force according to the protocol of collection of data. The response rate reflected by the teachers was very low, and finally, we got 214 questionnaire sheets including fully and partially filled.

After collecting the data, five-digit unique teachers' ID corresponding to each teacher was generated. Teacher's data was arranged into *Microsoft Excel* sheet and also stored in *Microsoft Access* to develop the database in future. For RC-network, the research data for each teacher is collected through *Google Scholar* on the Internet. This data is arranged in *Microsoft Excel* and processed and modified to feed into Cytoscape [18] as input data. The RC-network was drawn using this data.

For Comm-network, the data was fed into excel sheet through the questionnaire. Same five-digit unique ID was used to construct the Comm-network. Comm-network was also constructed using Cytoscape 3.4.0. After drawing the network in Cytoscape, we analyzed the topological, statistical, and descriptive properties of the RC-network and Comm-network such as degree, degree distribution, density, cluster coefficient, average path length, diameter, and centrality. A comparison between RC-network and Comm-network was done on the basis of these parameters. We used a Cytoscape app MCODE [19] to find the clusters (communities) in both RC-network and Comm-network.

3 Result and Discussion

Constructed RC-network having 712 nodes and 747 edges is shown in Fig. 1. Self-loops and multi-edges are removed during construction of network. There are seven isolated nodes in the RC-network. The connected components depict the connectivity of whole network, in addition a lower number of connected components ensure a high connectivity and vice versa. In RC-network, there are 146 connected components (Table 1). It suggests that RC-network does not have a strong connectivity. Characteristic path length for RC-network is 5.774, i.e., on the average the distance between two connected nodes is 5.774. RC-network is characterized by the local clusters formed by individuals who are linked to most of others.

The tendency of clustering is measured by clustering coefficient. The clustering coefficient of RC-network is 0.201. The diameter of the network is 15 which indicates the maximum distance between any two nodes in a component of the network. This signifies that the network is not a small world [16]. This is evident from the scarcity of connections and multiplicity of components. Average number of neighboring nodes of a node v shows average connectivity of node v in the network. In the RC-network, average number of neighbors is 2.098. The average connectivity in normalized version (value between 0 and 1) is the network density. In RC-network, the network density is 0.003 which shows that network is weakly populated with edges.

A network has perfect star topology when network centralization is 1 and decentralized network structure when network centralization is 0. Therefore, centralization is the measure of network centrality in some sense. The value of network centralization in RC-network is 0.015. RC-network contains 4% (20,664 in number) paths of the maximum possible shortest paths.

Node degree distribution for RC-network in Fig. 2a follows the power law which is fitted by the equation $y = ax^b$ with the values of $a = 530.03$ and $b = 2.187$.

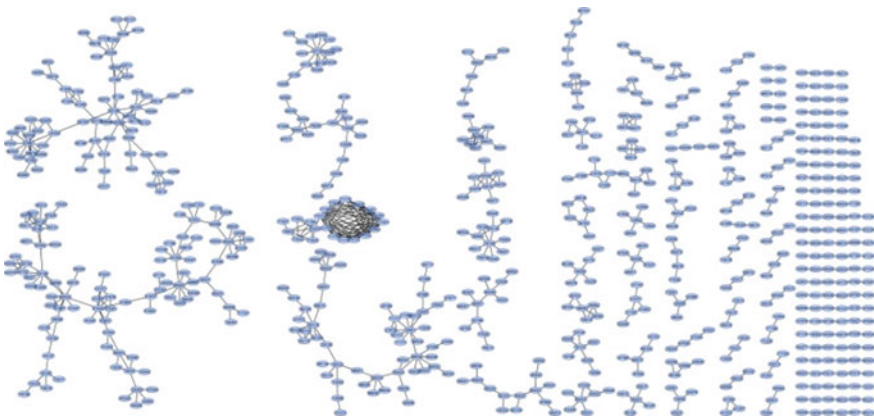


Fig. 1 RC-network

Table 1 Some parameters of RC-network and Comm-network

| Parameter | Value (RC-network) | Value (Comm-network) |
|-----------------------------|--------------------|----------------------|
| Clustering coefficient | 0.201 | 0.028 |
| Connected components | 146 | 105 |
| Network diameter | 15 | 18 |
| Network radius | 1 | 1 |
| Network centralization | 0.015 | 0.017 |
| Shortest paths | 20,664 (4%) | 74,818 (8%) |
| Characteristic path length | 5.774 | 6.718 |
| Average number of neighbors | 2.098 | 1.860 |
| Number of nodes | 712 | 943 |
| Network density | 0.003 | 0.002 |
| Network heterogeneity | 0.933 | 1.208 |
| Isolated nodes | 7 | 1 |
| Number of self-loops | 0 | 0 |
| Multi-edge node pairs | 0 | 0 |

Correlation coefficient is calculated 0.982. *R*-squared value on logarithm scale is computed as 0.875. In Fig. 2b, average clustering coefficient distribution for RC-network is shown. Network diameter and shortest path length distribution are used to indicate the small-world property of the network [16]. The frequency distribution of path lengths in Fig. 2c depicts that frequency of paths having lengths 6 is the highest and frequency having paths length 15 is the smallest. Topological coefficient measures the tendency of a vertex having shared neighbors in the network. In other words, it gives the relative measurement to what extent a vertex shares neighbors with other vertices. Network analyzer in Cytoscape measures the topological coefficient for each vertex having more than one neighbor in the network [20]. Topological coefficient distribution is shown in Fig. 2d depicting that there are fewer vertices whose topological coefficient is high.

Cytoscape app MCODE 1.4.2 [19] is used to find clusters in the RC-network. Modular Complex Detection (MCODE) is a tool to detect densely connected regions in the given network. We extracted a total of 38 clusters using Haircut method, and degree cutoff value is 2 (Table 2a). Clusters having more than five nodes are shown in Fig. 3a, b. RC-network has only 747 edges so it is not so dense as we expected. We got only 2–3 clusters in RC-network which are dense. These clusters are formed by the teachers of the same department or the same college/center indicating that research is not interdisciplinary in nature.

The constructed Comm-network using Cytoscape 3.4.0 is shown in Fig. 4. The Comm-network has 943 nodes, whereas there were 712 nodes in RC-network. Therefore, Comm-network has 231 more nodes with respect to RC-network, but network

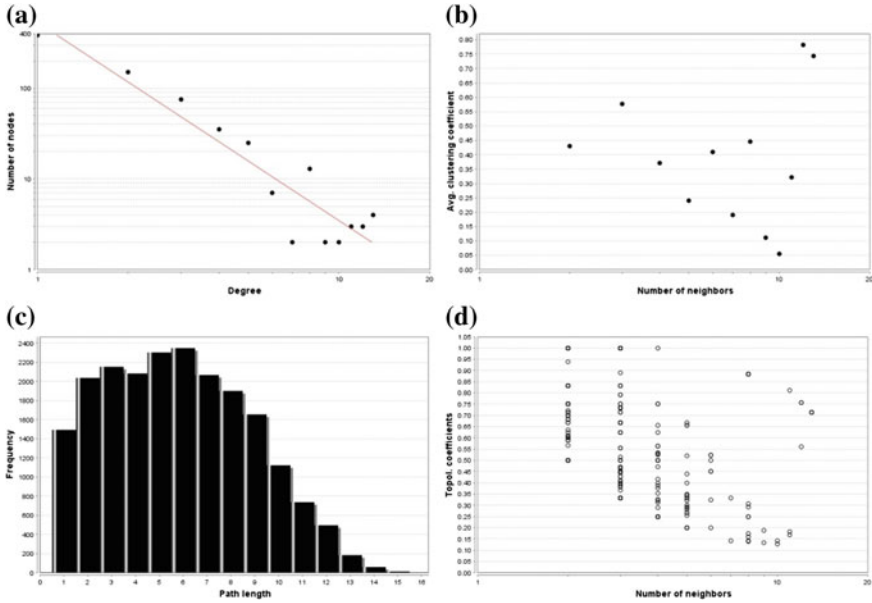


Fig. 2 **a** Node degree distribution of RC-network, **b** average clustering coefficient of RC-network, **c** path length frequency distribution of RC-network, and **d** topological coefficient distribution of RC-network

Table 2 (a) Clusters’ description in RC-network, and (b) clusters’ description in Comm-network

| S. No | Score | #nodes | #edges | #clusters |
|------------|--------|--------|--------|-----------|
| (a) | | | | |
| 1 | 11.167 | 13 | 67 | 1 |
| 2 | 5 | 5 | 10 | 1 |
| 3 | 4.8 | 6 | 12 | 1 |
| 4 | 4.5 | 5 | 9 | 3 |
| 5 | 4 | 4 | 6 | 9 |
| 6 | 3 | 3 | 3 | 23 |
| (b) | | | | |
| 1 | 2.8 | 6 | 7 | 1 |
| 2 | 2.737 | 20 | 26 | 1 |
| 3 | 2.333 | 13 | 14 | 1 |
| 4 | 2.333 | 7 | 7 | 1 |
| 5 | 2.19 | 22 | 23 | 1 |
| 6 | 2.167 | 13 | 13 | 1 |
| 7 | 2.154 | 14 | 14 | 1 |
| 8 | 2.095 | 22 | 22 | 1 |

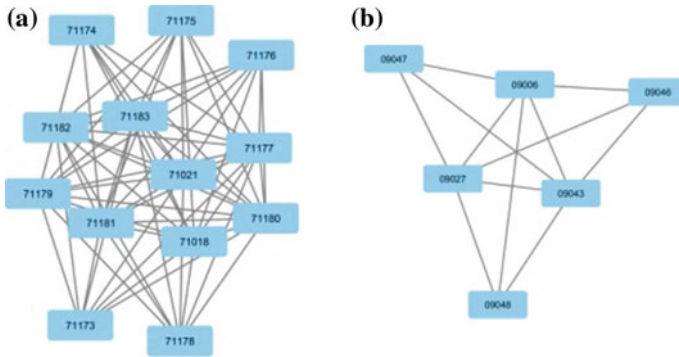


Fig. 3 a Cluster of RC-network with 11.67 score having 13 nodes and 67 edges, b Cluster of RC-network with 4.8 score having 6 nodes and 12 edges

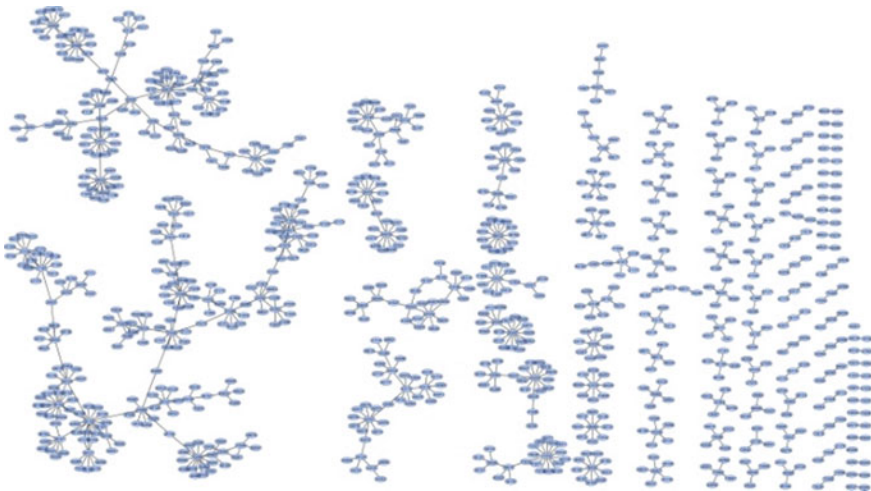


Fig. 4 Comm-network

density for Comm-network is 0.002 which is less than 0.003 the density of RC-network.

Number of isolated node in Comm-network is 1, which may be due to errors in collecting data, whereas there were seven isolated nodes in RC-network. Number of self-loops and multi-edges node pairs are zero in Comm-network same as in the RC-network. The value of clustering coefficient for Comm-network is 0.028 which is near about to the clustering coefficient of RC-network.

The numbers of connected components are 105 in Comm-network which are less than 146 (number of connected components in RC-network). Network diameter of Comm-network is 18 which is greater than the diameter of RC-network. Network centralization of Comm-network is 0.017 which is almost same as of RC-network.



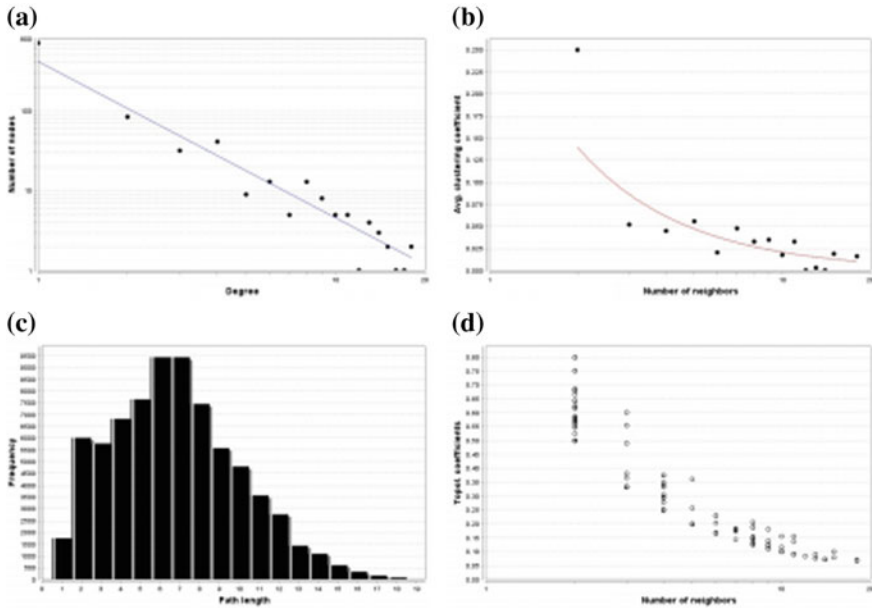


Fig. 5 **a** Node degree distribution of Comm-network, **b** average clustering coefficient distribution of Comm-network, **c** path length frequency distribution of Comm-network, and **d** topological coefficient distribution of Comm-network

Number of shortest paths is 8% (74,818 in numbers) in Comm-network, a little higher than the number of paths (4%) in the RC-network.

The characteristic path length of Comm-network is 6.718, whereas it was 5.774 in the RC-network. The average number of neighbors in Comm-network is 1.860 which is less than 2.098 (the average numbers of neighbors in RC-network) (Table 1).

Node degree distribution for Comm-network is fitted with power law by the equation $y = ax^b$ with the values of $a = 420.44$ and $b = 1.961$ (Fig. 5a). The correlation value and R -squared values for node degree distribution of Comm-network are computed. The R -squared value 0.907 in the case of Comm-network is high in comparison to 0.875 in the case of RC-network. Therefore, node degree distribution in Comm-network is more accurately fitted in comparison to RC-network. Average cluster coefficient distribution is also fitted by power law by the equation $y = ax^b$ with parameter values $a = 0.316$ and $b = -1.182$ (Fig. 5b). Correlation coefficient and R -squared are also computed whose values are 0.889 and 0.603, respectively. R -squared value is computed on logarithm scale. The frequency distribution of path lengths in Comm-network (Fig. 5c) shows that frequencies of paths with lengths 6 and 7 are highest.

The range of path lengths in Comm-network is 1–18, whereas it is 1–15 in RC-network. The number of paths having length 18 is the smallest in Comm-network. The topological coefficient distribution in Fig. 5d suggests that number of nodes is

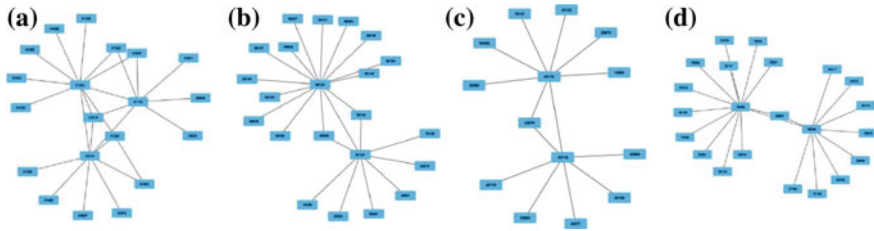


Fig. 6 **a** Cluster with 2.737 score having 20 nodes and 26 edges, **b** cluster with 2.19 score having 22 nodes and 23 edges, **c** cluster with 2.154 score having 14 nodes and 14 edges, and **d** cluster with 2.085 score having 22 nodes and 22 edges

less having higher topological coefficient same as in the RC-network. Totally eight clusters are extracted in Comm-network using the MCODE 1.4.2. The description about all these eight clusters is available in Table 2b. The visualization of clusters having more than 13 nodes is shown in Fig. 6a–d.

4 Conclusion

The present study may be useful for the administration of Delhi University to find out the tendency of collaboration of each teacher on different parameters like college, department, subjects, etc. This study is also useful for identifying the emerging research areas for research students. In future, proposed study may be generalized or applied to other universities in creating the database of researchers, teachers, or scientists to develop a common platform for collaboration and good-quality research. This study may also be helpful to policy-makers in terms of economy to release the grants to different colleges, department/centers, or groups of teachers.

Acknowledgements This work was carried out under the innovation project 2015-16 funded by University of Delhi. We acknowledge University of Delhi for providing the fund for this project. We also acknowledge the students Aanchal Chawla, Aditi Singh, Avni Bhasin, Masoom Bhargava, Monica Pruthi, Parina Rattan, Parul Jain, Sagarika Seth, Shivani Dayal, Vertika Shukla, and Jyoti Dayal for collecting the data.

References

1. Braun, T. (2005). Handbook of Quantitative Science and Technology Research. The Use of Publication and Patent Statistics in Studies of S&T Systems. *Scientometrics*, 63(1), 185–188.
2. Freeman, C., & Soete, L. (2009). Developing science, technology and innovation indicators: What we can learn from the past. *Research Policy*, 38(4), 583–589.
3. Leieken, S., & Kempner, R. (2010). Collaborate leading regional innovation clusters. *The Council on Competitiveness*.

4. Beaver, D. D. (2001). Reflections on scientific collaboration (and its study): Past, present, and future. *Scientometrics*, 52(3), 365–377.
5. Bozeman, B., & Corley, E. (2004). Scientists' collaboration strategies: Implications for scientific and technical human capital. *Research Policy*, 33(4), 599–616.
6. Katz, J. S., & Martin, B. R. (1997). What is research collaboration? *Research Policy*, 26(1), 1–18.
7. Borgman, C. L., & Furner, J. (2002). Scholarly communication and bibliometrics. *Annual Review of Information Science and Technology*, 36(1), 2–72.
8. Melin, G. (2000). Pragmatism and self-organization: Research collaboration on the individual level. *Research Policy*, 29(1), 31–40.
9. Melin, G., & Persson, O. (1996). Studying research collaboration using co-authorships. *Scientometrics*, 36(3), 363–377.
10. Batagelj, V., & Mrvar, A. (2000). Some analyses of Erdos collaboration graph. *Social networks*, 22(2), 173–186.
11. Beaver, D., & Rosen, R. (1979). Studies in scientific collaboration Part III. Professionalization and the natural history of modern scientific co-authorship. *Scientometrics*, 1(3), 231–245.
12. Egghe, L., Rousseau, R., & Van Hooydonk, G. (2000). Methods for accrediting publications to authors or countries: Consequences for evaluation studies. *Journal of the American Society for Information Science*, 51(2), 145–157.
13. Barabási, A. L., Jeong, H., Néda, Z., Ravasz, E., Schubert, A., & Vicsek, T. (2002). Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and its Applications*, 311(3–4), 590–614.
14. Newman, M. E. (2001). Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. *Physical Review E*, 64(1), 016132.
15. Newman, M. E. J. (2001). The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, 98(2), 404–409.
16. Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, 393(6684), 440–442.
17. Hagstrom, W. O. (1965). *The scientific community*. Carbondale: Southern Illinois University Press.
18. Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., ... & Ideker, T. (2003). Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13(11), 2498–2504.
19. Bader, G. D., & Hogue, C. W. (2003, January). An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, 4(1), 2.
20. Stelzl, U., Worm, U., Lalowski, M., Haenig, C., Brembeck, F. H., Goehler, H., ... & Timm, J. (2005). A human protein-protein interaction network: A resource for annotating the proteome. *Cell*, 122(6), 957–968.