Advances in Intelligent Systems and Computing 413

H.S. Saini
Rishi Sayal
Sandeep Singh Rawat *Editors*

Innovations in Computer Science and Engineering

Proceedings of the Third ICICSE, 2015

Springer

Advances in Intelligent Systems and Computing

Volume 413

Series editor

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

About this Series

The series “Advances in Intelligent Systems and Computing” contains publications on theory, applications, and design methods of Intelligent Systems and Intelligent Computing. Virtually all disciplines such as engineering, natural sciences, computer and information science, ICT, economics, business, e-commerce, environment, healthcare, life science are covered. The list of topics spans all the areas of modern intelligent systems and computing.

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

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

H.S. Saini · Rishi Sayal · Sandeep Singh Rawat
Editors

Innovations in Computer Science and Engineering

Proceedings of the Third ICICSE, 2015

Springer

Editors

H.S. Saini
Guru Nanak Institutions
Ibrahimpatnam, Telangana
India

Sandeep Singh Rawat
Guru Nanak Institutions
Ibrahimpatnam, Telangana
India

Rishi Sayal
Guru Nanak Institutions
Ibrahimpatnam, Telangana
India

ISSN 2194-5357 ISSN 2194-5365 (electronic)
Advances in Intelligent Systems and Computing
ISBN 978-981-10-0417-9 ISBN 978-981-10-0419-3 (eBook)
DOI 10.1007/978-981-10-0419-3

Library of Congress Control Number: 2015960405

Printed on acid-free paper

This Springer imprint is published by SpringerNature
The registered company is Springer Science+Business Media Singapore Pte Ltd.

Preface

The volume contains 36 papers presented at the Third International Conference on Innovations in Computer Science and Engineering (ICICSE 2015) held during 7– 8 August, 2015 at Guru Nanak Institutions campus in association with CSI Hyderabad Chapter.

The focus of the 3rd ICICSE 2015 is to provide an opportunity for all professionals and aspiring researchers, scientists, academicians, and engineers to exchange their innovative ideas and new research findings in the field of computer science and engineering. We have taken an innovative approach to give an enhanced platform for these personnel, participants, researchers, students, and other distinguished delegates to share their research expertise, experiment breakthroughs, or vision in a broad criterion of several emerging aspects of the computing industry. It received a good number of submissions from different areas related to innovation in the field of computer science. After a rigorous peer-review process with the help of our program committee members and external reviewers, we finally accepted 36 submissions with an acceptance ratio of 0.26.

ICICSE 2015 along with DECODE IT-Park was inaugurated by the Hon'ble Minister IT & Panchayat Raj, Shri K. Taraka Rama Rao, and the Plenary Session of the ICICSE 2015 was conducted by Mr. Lloyd Sanford, CEO, Top Blue Logistics.

We take this opportunity to thank all the keynote speakers and Special Session Chairs for their excellent support to make ICICSE 2015 a grand success. We would like to thank all reviewers for their time and effort in reviewing the papers. Without this commitment it would not have been possible to have the important 'referee' status assigned to papers in the proceedings. The quality of these papers is a tribute to the authors and also to the reviewers who have guided any necessary improvement. We are indebted to the program committee members and external reviewers who not only produced excellent reviews but also did in short time frames. We also thank CSI Hyderabad Chapter for coming forward to support us to organize this mega event.

We thank the authors and participants of this conference. Special thanks to all the volunteers, without whose tireless efforts we could not arrange to run the conference smoothly. All the efforts are worthy and it would please us all if the readers of this proceedings and the participants of this conference found the papers and event inspiring and enjoyable.

Finally, we place our special sincere thanks to the press, print, and electronic media for their excellent coverage of this conference.

H.S. Saini
Rishi Sayal
Sandeep Singh Rawat

Committee

Patrons

Sardar Tavinder Singh Kohli
Sardar Gagandeep Singh Kohli

Conference Chair

Dr. H.S. Saini

Conference Co-chairs

Dr. Veeranna
Dr. D.D. Sharma
Dr. S. Sreenatha Reddy
Dr. Rishi Sayal

Convenors

Dr. S. Masood Ahamed
Prof. V. Deva Sekhar
Dr. Sandeep Singh Rawat

Co-convenors

Dr. Vijayalakshmi
Prof. Dathatreya
Dr. V. Sathiyasuntharam
Mr. Lalu Nayak
Prof. N. Prasanna Balaji
Dr. Aniruddha Bhattacharjya
Mr. S. Madhu
Mrs. Subbalakshmi

Conference Committee

Dr. Rishi Sayal
Ms. Tayyaba Khatoon
Ms. P. Harsha
Mr. B. Nandan
Mr. Manikrao Patil

Publicity Chair International

Dr. D.D. Sharma
Mr. Imran Quereshi
Ms. Kanchanlatha
Dr. Aniruddha Bhattacharjya
Mr. V. Poorna Chandra

Publicity Chair National

Prof. V. Deva Sekhar
Ms. D. Sirisha
Ms. B. Mamatha
Mr. D. Saidulu
Mr. Y. Ravi Kumar

Program and Publication Chair

Dr. Sandeep Singh Rawat
Mr. T. Ravindra
Mrs. K. Prasuna
Mr. K. Suresh
Mr. Devi Prasad Mishra
Mr. Nusrat Khan

Accommodation Committee

Dr. S. Masood Ahamed
Mr. A. Ravi
Mr. Vinay Sagar
Mr. B. Sudhakar
Mr. A. Srinivas

Advisory Board—International/National, Technical Program Committee

Dr. Robin Doss, Australia
Dr. San Murugesan, Australia
Dr. Chandrashekar Commuri
Yang, Lung-Jieh, Taiwan
Dr. William Oakes, USA
Dr. Sartaj Sahni, USA
Dr. Jun Suzuki, USA

Dr. Prabhat Kumar Mahanti, Canada
Mrs. B. Sunitha, Melbourne, Australiax
M. Siva Ganesh, USA
Dr. Muzammil H Mohammed, Saudi Arabia
Dr. Raj Kamal, India
Dr. A. Govardhan, India
Dr. Naveen Kumar, India
Dr. Uday Bhaskar Vemulapati, India
Dr. A. Sadanandam, India
Dr. R.B.V. Subramanyam, India
Prof. S.V. Raghavan, India
Dr. Durgesh Kumar Mishra, India
Mr. Krishan Murthy, India
Dr. Sohan Garg, India
Dr. Zubair Baig, Australia
Dr. Hemant Pendharkar, USA
Dr. Sitalakshmi Venkatraman, Australia
Dr. Muzammil H Mohammed, Saudi Arabia
Dr. P.S. Grover, India
Dr. B. Anuradha, India
Dr. A.V.N. Krishna, India
Dr. V.V.S.S.S. Balaram, India
Mr. Michael W. Osborn
Prof. Rajkumar Buyya, Australia
Dr. S. Jimmy Gandhi, USA
Dr. Vinod Lohani, USA
Dr. Stephanie Farell, USA
Dr. Arun Somani, USA
Prof. Pascal Lorenz, France
Dr. Vamsi Chowdavaram, Canada
Mr. M. Kiran, CTS, New Jersey, USA
Dr. Lakshmivarahan, USA
Dr. S.R. Subramanya, USA
Dr. V. Vijay Kumar, India
Dr. Aruna Malapadi, India
Mr. Y. Prasad, India
Mr. Ravi Sathanapalli, India
Dr. Somayajulu, India
Dr. P.S. Avadani, India
Mr. H.R. Mohan, India
Prof. Atul Negi, India
Dr. R. Sridevi, India
Mr. Hari Devarapalli, India
Dr. Lakshmivarahan, USA
Prof. Soura Dasgupta, USA

Prof. San Murugesan, Australia
Dr. Lalit Awasthy, India
Dr. Aruna Malapadi, India
Dr. Nagu Chandra Sekhar Reddy, India
Dr. K.V.G. Rao, India
Dr. M. Varaprasad Rao, India

A Note from the Organizing Committee

Welcome to the 3rd International Conference on Innovations in Computer Science and Engineering, India. On behalf of the entire organizing committee, we are pleased to welcome you to ICICSE 2015.

ICICSE, as the conference in the field, offers a diverse program of research, education, and practice-oriented content that will engage computer science engineers from around the world. The two-day core of the meeting is anchored by the research paper track. This year, the research paper track received 181 submissions. The papers underwent a rigorous two-phase peer review process, with at least two Program Committee members reviewing each paper. The Program Committee selected 36 papers. All members of the Program Committee attended the meeting. These papers represent world-wide research results in computer science engineering.

Planning and overseeing the execution of a meeting of ICICSE is an enormous undertaking. Making ICICSE 2015 happen involved the combined labor of more than 50 volunteers contributing a tremendous amount of time and effort. We offer our sincere thanks to all of the committee members and volunteers, and encourage you to take the opportunity to thank them if you meet them at the conference. We also thank all our sponsors who helped to make this event accessible to the computer science engineering community.

Finally, we thank the editorial board of *Springer* for agreeing to publish the proceedings in *Springer* and the staff at the editorial office for all their help in the preparation of the Proceedings.

Dr. H.S. Saini
Professor & Managing Director

Dr. Rishi Sayal
Dean (Academic & Training)

Dr. Sandeep Singh Rawat
Professor and Head—CSE and IT

Contents

About the Editors

Dr. H.S. Saini Managing Director of Guru Nanak Institutions obtained his Ph.D. in the field of Computer Science. He has over 22 years of experience at the university/college level in teaching UG/PG students and has guided several B. Tech., M.Tech., and Ph.D. projects. He has published/presented high quality research papers in international, national journals and proceedings of international conferences. He has two books to his credit. Dr. Saini is a lover of innovation and is an advisor for NBA/NAAC accreditation process to many institutions in India and abroad.

Dr. Rishi Sayal Dean, Academics and Training—Guru Nanak Institutions Technical Campus has done B.E. (CSE), M.Tech (IT), Ph.D. (CSE), LMCSI, LMISTE, MIEEE, and MIAENG (USA). He completed his Ph.D. in Computer Science and Engineering in the field of Data Mining from the prestigious and oldest Mysore University of Karnataka state. He has over 23 years of experience in training, consultancy, teaching, and placements. His current areas of research interest include data mining, network security, and databases.

Dr. Sandeep Singh Rawat Professor and Head—CSE and IT, obtained his Bachelor of Engineering in Computer Science from National Institute of Technology, Surat (formerly REC Surat) and his Master's in Information Technology from Indian Institute of Technology, Roorkee. He has been awarded a doctorate in Computer Science and Engineering by University College of Engineering, Osmania University, Hyderabad 2014. He is working at Guru Nanak Institutions Hyderabad since 2009. He has 12 years of teaching and 2 years of industrial experience. His current research interests include data mining, grid computing, and data warehouse technologies. He is a Life Member of technical societies like CSI and ISTE and a member of IEEE.

Secure ATM Door Locking System Using RFID

Sandeep Singh Rawat, Shaik Saidulu and Rasmi Ranjan

Abstract The recent incident at one of the ATMs in Bengaluru has challenged the security and safety measures currently used at ATM machines. Anyone can enter the ATM cabin without the knowledge or permission of the person currently using the ATM. This increases the possibility of crime-like attacks and theft at gun points, etc., as another customer can easily enter the ATM at any time. The existing system does not have measures to ensure that when a person is using the ATM, no other person is allowed to enter. The proposed system ensures that no other person is allowed to enter the ATM cabin when a person is using the ATM machine. This system will have an electronic lock that has to be unlocked by swiping the person's debit card, and once the person has entered the cabin the door automatically locks once it is closed. Now, to open the door the same debit card has to be used from inside the cabin. In this way, a person who is waiting outside has no chance of entering the cabin and attack the person who is currently using the ATM. This system not only ensures safety against attacks and thefts but also creates awareness against antisocial elements. After developing this application a fair amount of revenue can be generated by collaborating with banks and the IT industry.

Keywords ATM · Data analysis · Security · RFID

S.S. Rawat (✉) · S. Saidulu · R. Ranjan
Guru Nanak Institute of Technology, Hyderabad, India
e-mail: isandeepi@yahoo.com

S. Saidulu
e-mail: sk.saidulu@gmail.com

R. Ranjan
e-mail: ranjandandpat@gmail.com

H.S. Saini et al. (eds.), *Innovations in Computer Science and Engineering*,
Advances in Intelligent Systems and Computing 413,
DOI 10.1007/978-981-10-0419-3_1

1 Introduction

Secure ATM door locking system using RFID is a system to secure ATM doors so that it is more secure and provides better security to users. Today's ATM systems allow users to enter the ATM at any time without much security or reliability. Any person possessing a card similar to an ATM, debit,or credit card is allowed to enter the ATM. There is no prior security to check whether the person using the ATM has a valid card or not. The person trying to withdraw money is not sure if somebody is looking over or entering the ATM. The current system does not restrict users to enter ATM machines in case there is already a user inside. This project mainly focuses on providing security to all customers who come to an ATM system.

Current ATM systems have a card scanner that scans the customer card and then lets them inside. But this system is not efficient and does not work properly. If a person is indoors, the system does not restrict another user from opening the door. This can be dangerous sometimes. There is no security for the user inside the ATM. Sometimes this can also be life-threatening to the user. It is dangerous especially in areas where there is not much public movement or it is a lonely area. At these locations it is easy to attack a person using the ATM and flee with the money.

Thefts are also prominent in areas of low density population. Robbers simply break open the door and steal the money. They may use afake card to get entry to the ATM machine. Installing CCTV cameras and even having a security guard at the door does not help. The security camera can be easily destroyed and security can be breached. Installation of CCTV cameras and having a security guard at the ATM is costly. If a person faints inside the ATM room there is no way to detect that and inform others. Sometimes, nobody notices this for at least a period of an hour. By this time the person might reach a critical condition. So we need to make sure that people entering the ATM must leave within a certain amount of time, else some alert must be sent to the nearby authorities. This would also prevent robbery, as they would not have much time, and if they take a lot of time then an alarm will be raised and notified to the nearby authorities. This work can be used in the ATM system to better secure the ATM machines. It can also be used in companies where there is need for restriction to a room. Every company might have some confidential information and in order to secure that they generally have a separate room. This system can be used to secure such rooms, especially server rooms. The system can be modified to be used in educational systems to track the number of students entering the premises and the time at which they are leaving.

The rest of the paper is organized as follows. Section 2 discusses the related work. Section 3 explains the solution framework used for secure ATM machines. Section 4 gives some reflections about the observed results. Finally, Sect. 5 concludes with mention of the future likely enhancements of the system.

2 Related Work

RFID technology, which stands for Radio Frequency Identification, is a popular and widely used technology in the world. Using electromagnetic waves to transfer data, this technology is used to automatically track or identify objects or persons who pose an RFID chip/tag/card, mostly in automated industries. Apart from commercial usage, RFID is also used for domestic purposes like in shopping malls to prevent thefts. Commercially available RFID devices operate in the standard ISM frequency band of 13.56 MHz Usually, the RFID chip embedded into an object transmits a unique identification number received by the RFID reader which makes the object detectable [1].

To prevent thefts, attacks on customers, security measures at banks play a crucial role. Automated teller machines being a quick money withdrawing service are the prime target of criminals to get easy money. In order to ensure a safe and secure banking environment for customers, it is necessary for a flawless security system to be employed in banking services, especially for ATMs. Magnetic strip smart cards which are now being used to identify customers, are highly vulnerable for malpractices such as ATM scams. Scammers place a transparent, slim, rigid plastic sleeve-like cover into the ATM and wait for their victim. When the customer slips his card into the ATM, it gets stuck inside and being unable to identify the card, it asks to use the access pin number multiple times, rejecting each entry assuming to be wrong. Eventually the customer leaves with the belief that his card is blocked. Thieves now use small prongs to pull out the sleeve along with the ATM access card inside it. With this card, they empty the customer's account using the pin they obtain by peeking in while the customer is typing. When considering vulnerabilities and causation in civil lawsuit, extreme measures of security are of supreme importance. A biometric measure as a means of enhancing the security has emerged from discourse. The primary focus of this work is to develop a biometric strategy to enhance the security features of the ATM for effective banking transaction [2].

The importance of integrating the fingerprint of the user into the bank's database to further authenticate has been discussed by Prof Selina Oko and Jane Oruh in their paper to improve the existing security system of ATMs. Fingerprint technology can provide a much more reliable and accurate user authentication method. Various methods of frauds like ATM card theft, pin theft, skimming, card block method, pin pad method, force withdrawals, and many more can be prevented. Based on fingerprint GSM technology, there is a system used to implement high security ATMs and locker systems that can be established in banks and secured offices. In this system money can only be recovered by the authentic person. Banks will collect the customer fingerprints and mobile phone number while opening the account and only this person will have access to the ATM to draw money from a certain account. The working of this system can be observed when the customer places his finger on the fingerprint scanning module of the ATM. After identifying account from the fingerprint, the system generates a 4-digit one-time password which is sent to the registered mobile using which the customer can draw money. To implement high security locker systems, there is another system based on RFID, fingerprint, and GSM technology which can be implemented in banks, offices, and homes. Only the authentic person can recover money/belongings from the locker using this system. Using this system based on RFID, fingerprint, and GSM technology, we can implement the automatic door locking system which can initiate, authenticate, and validate the user and unlock the door in real-time for locker secure access [3].

Most of the previous works assume the communication channel between an RFID reader and its backend server is secure and concentrate only on the security enhancement between the RFID tag and RFID reader. However, once the RFID reader modules get immensely deployed in the consumer's handheld devices, the privacy violation problems at the reader side will become a matter of concern for individuals and organizations. If the future communication environment for RFID systems is to be wireless, it increases insecurity among the three roles. We need to achieve message security, anonymity, availability, and protection of information from being stolen or tampered with. Recently, the use of mobile devices has become commonplace worldwide. They have the functionality to read RFID tags and they also have higher computing potential. During the transaction process they take less time for encryption, decryption, and certification. RFID formed smart stick prototype have been developed to aid and assist the visually challenged (user) in shopping through GORE (Goal Oriented Requirements Engineering methodology) [4].

A system called campus security system, to make the college campus secure in every way needs to be done and also to maintain discipline in the education campus in this way by reducing the loudness of horns is developed. The proper functions are to keep a log of the person entering the campus automatically. Only RFID installed vehicles can enter in college campus. The automatic vehicle tracking facility delivers flexibility, suitability, and responsiveness [5]. Identification by radio frequencies in health care is one of its major growth areas. This system describes how this latest technologies are used to build a smart hospital. It also shows how to use an assets tracking application, to enhance the quality of the health care services [6]. Prof. Selina Oko and Jane Oruh discussed in their paper, the existing security of the Automated Teller Machine has been improved by integrating the fingerprint of the user into the database as to further authenticate it. Fingerprint technology can provide a much more accurate and reliable user authentication method [7]. Most of the previous works assume the communication channel between an RFID reader and its database server is secure and concentrates only on the security enhancement between the RFID tag and reader. However, once RFID reader modules gets deployed in consumers' devices (mostly mobile), the privacy violation problems at reader side will become a major concern for any organizations [8]. A system called campus security system, to make the college campus secured in very way that is need to be done and also maintaining the discipline in the educational campus by reducing the loudness of the horn, is developed. The main functions are to keep a log of the every vehicle coming to the organization automatically. This automatic vehicle tracking system gives the better responsiveness for different organizations [9].

3 Proposed Approach

The complete system is divided into two modules. They are:

- Card Reader
- ATM application

3.1 Card Reader Module

This module reads the card information and then verifies if the card being used is valid ATM application or not. Here the application is used to simulate the same functionality as the ATM and makes sure that the transaction is carried out smoothly. It also checks the time for the transaction and in case it is too high it automatically alerts the authorities.

In this system, (Fig. 1) we use 5 V power supply for the microcontroller. We use rectifiers for converting A.C to D.C and a step-down transformer to step down the voltage. Microcontrollers were originally used as components in process-control systems. However, due to their small size and low price, microcontrollers are now also used in regulators for individual control loops. In several areas microcontrollers are now outperforming the analog counterparts and are cheaper. The microcontroller used here is the Microcontroller AT89S52 belonging to the 8051 family.

RFID is the wireless noncontact use of radio-frequencies to transfer data for the purposes of involuntarily identifying and tracking tags attached to objects. The tags contain electronically preserved data. Unlike a barcode, the tag does not undoubtedly need to be within line of sight of the interpreter, and may be embedded in the tracked object. The use of RFID here is to enhance the automation of the attendance using the RFID reader module and the RFID tags. LCD block is used to know the status of the devices or operation. It is used to display student details such as student's name, login time, and logout time. MAX232 is an IC that converts signals from an RS-232 serial port to signals suitable for use in TTL compatible

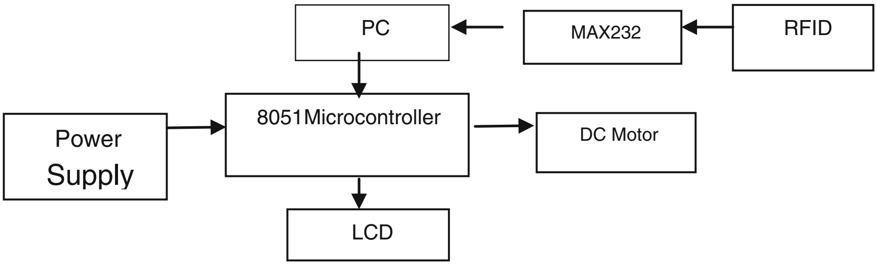


Fig. 1 Block diagram

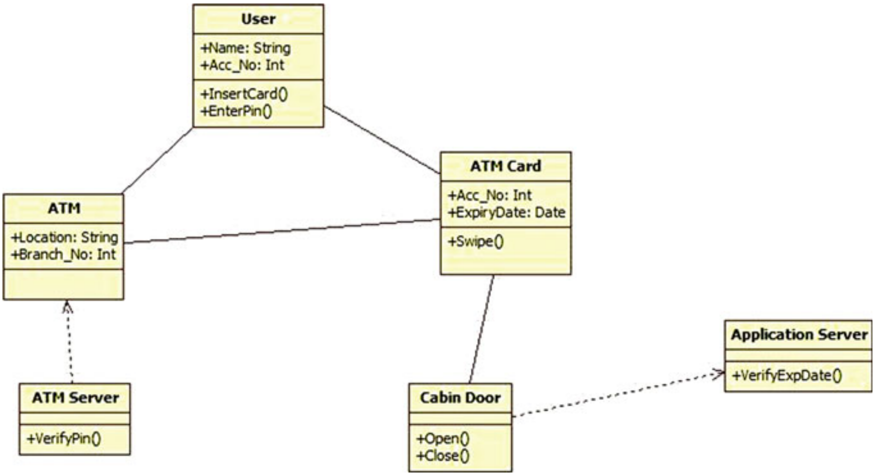


Fig. 2 Class diagram of proposed system

digital logic circuits. MAX232 is a twofold driver/receiver and typically translates the RX, TX, CTS, and RTS signals.

In Fig. 2, user, ATM, ATM card, ATM server, cabin door, and application server are the different classes. The ATM has two attributes, location and branch number. The ATM card has two attributes, account number and expiry date, and an operation, swipe. The ATM server's operation is only to verify the pin number that was entered in the ATM. The cabin door can be opened and closed according to the person entering into the ATM. The application server verifies if the ATM card is valid.

4 Experimental Results

We have simulated the ATM transaction process to provide a secure door using RFID. In our experimental tests we used dot net and SQL server in Windows environment. Figure 3a shows the form when it is waiting for the input to be entered. The input is the information read from the RFID reader. The RFID reader scans the card and finally transmits it by serial communication and the information is displayed on the screen. The card is then validated and the door is opened. Figure 3b shows the pin being entered. The pin submitted is echoed as stars so that it is impossible for others to see. If the valid pin is entered only then the person is allowed to proceed for further transaction. The pin entered must be a valid one, else it will be rejected. Figure 3c shows the form that can be used either to withdraw or check the balance of the customer. It is a simulation module and hence only two functionalities are added. The withdraw money option allows a user to draw money from the savings account. The user is given the functionality of pressing a key to choose an option. Option 'W' allows the user to withdraw the money and option

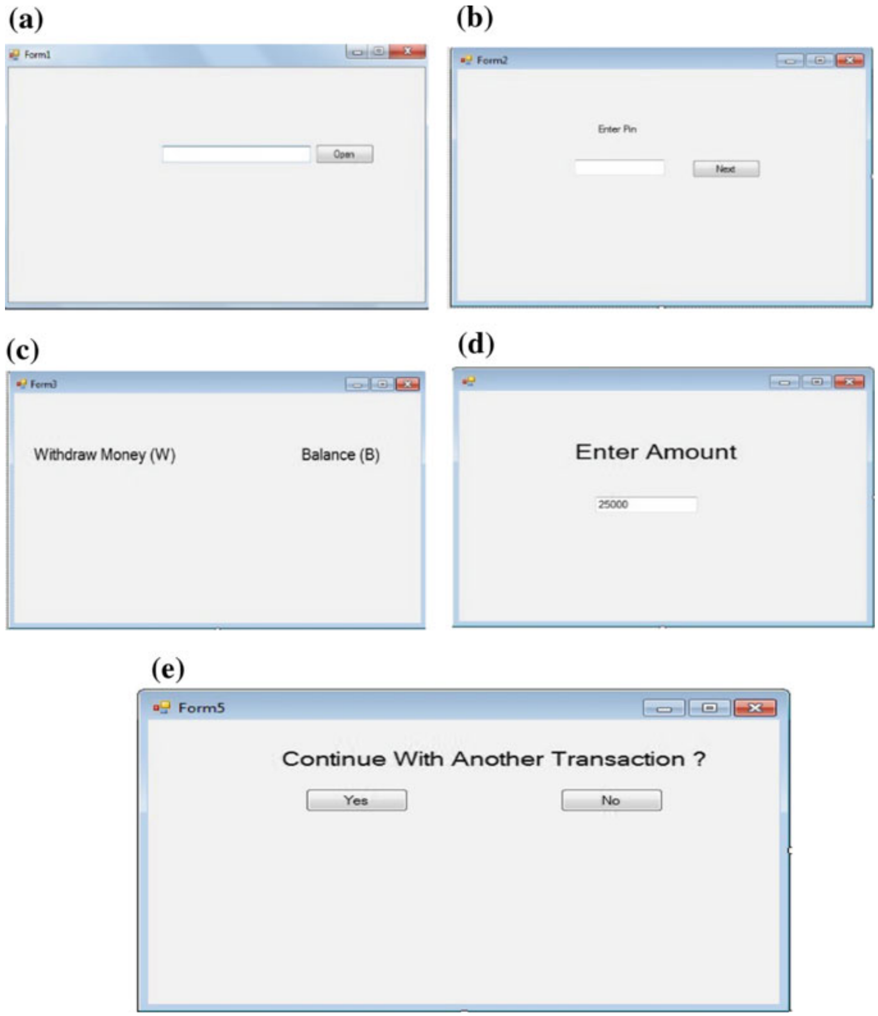


Fig. 3 Simulation of ATM transaction, **a** card reading, **b** entering pin, **c** check balance or withdraw, **d** entering amount to withdraw and **e** option to continue for another transaction

'B' allows the user to check his balance. Figure 3d allows the user to enter the amount in case he chooses the withdraw option on the previous screen. The amount once entered and accepted is deducted from his account immediately. Figure 3e shows the option of whether the customer wants to make another transaction. If he chooses yes, then a form for withdrawal opens again, else the initial form is opened for another card input. If the option chosen is 'no', then the door opens and another customer can now continue. A particular time is set. If the customer takes longer than that time, an alarm will be raised to notify the nearby person. Through this we can also prevent theft.

5 Conclusion and Future Work

This work helps to make the use of ATM systems much more secure. It also helps to detect any threats to the life of a person. People will be more careful and we can avoid a lot of thefts. The ATM system can be made much more secure. This system also reduces the cost of installation of CCTV cameras and security guards at the ATM machines.

Many directions for future enhancements are open. Among them, we can mention:

- The testing of the proposed work on a real environment and
- voice-based identification system to open and close the door.

References

1. Pramila D. Kamble, Dr. Bharti, W. Gawali, "Fingerprint Verification of ATM Security System by Using Biometric and Hybridization", IJSRP, Volume 2, Issue 11, November 2012.
2. Raghu Ram Gangi, Subhramanya Sarma Gollapudi, "Locker Opening And Closing System Using RFID, Fingerprint, Password And GSM".
3. Pennam Krishnamurthy & M. Maddhusudhan Redddy, "Implementation of ATM Security by Using Fingerprint recognition and GSM" International Journal of Electronics Communication and Computer Engineering Volume 3, Issue (1) NCRTCST, ISSN 2249–071X, 2012.
4. Ibidapo, O. Akinyemi, Zaccheous O. Omogbadegun, and Olufemi M. Oyelami, "Towards Designing a Biometric Measure for Enhancing ATM Security in Nigeria EBanking System".
5. Alexander De Luca, Marc Langheinrich, Heinrich Hussmann, "Towards Understanding ATM Security—A Field Study of Real World ATM Use".
6. Patrik Fuhrer and Dominique Guinard, "Building a Smart Hospital using RFID technologies".
7. Prof. Selina Oko and Jane Oruh, "Enhanced Atm Security System Using Biometrics".
8. Kopparapu Srivatsa, Madamshetti Yashwanth and A.Parvathy, "RFID & Mobile Fusion for Authenticated ATM Transaction".
9. Sushil I. Bakhtar, Dr. P.V. Ingole and Prof. Ram S. Dhekekar, "Design of RFID based Security System for College Campus".

Analysis of Various Parameters for Link Adaptation in Wireless Transmission

R.G. Purandare, S.P. Kshirsagar and S.M. Koli

Abstract Choosing a correct parameter from RSSI, SINR, PDR, and BER to estimate status of the wireless link is of paramount importance for link adaptation. RSSI may not point out interference effectively, whereas SINR thresholds are hardware dependant and require calibration. PDR and BER are measurable metrics with many considerations. It may project wrong status of link quality if not analyzed properly. Since reception of erroneous packet is frequent in wireless domain, cause has to be pinpointed accurately for pertinent remedial action. But two different causes require two different corrective actions, exponential back off for collisions, and change of transmission data rate or power for signal attenuation.

Keywords Channel state information matrix · Packet loss differentiation · Link adaptation

1 Introduction

Interference, collisions, and fading are inherent characteristics of wireless channel making the faithful delivery of data to the receiver difficult [1]. To improve the robustness and reliability of wireless transmissions, link adaptation techniques are

R.G. Purandare (✉)
Department of Electronics and Telecommunications,
Vishwakarma Institute of Information Technology, Pune, Maharashtra, India
e-mail: radhika.purandare@viit.ac.in

S.P. Kshirsagar
Anchortek Techno Consultancy Pvt. Ltd., Pune, Maharashtra, India
e-mail: shirish@anchorteksys.com

S.M. Koli
Department of Electronics and Telecommunications,
SKN College of Engineering, Pune, Maharashtra, India
e-mail: sanjaykoli@yahoo.com

H.S. Saini et al. (eds.), *Innovations in Computer Science and Engineering*,
Advances in Intelligent Systems and Computing 413,
DOI 10.1007/978-981-10-0419-3_2

employed. The effectiveness of which is dependent on true estimation and feedback of channel state information conveyed by the receiver to the transmitter.

The analysis in this paper covers rate adaptation. Earlier systems like Automatic Rate Fallback (ARF) used this without differentiating between signal fading and collisions. ARF uses the statistics of transmission success or failure. Adaptive ARF (AARF) [2] updates the thresholds for using higher or lower data rate. Automatic Rate Fallback (ARF) Collisions Aware Rate Adaptation (CARA) [3] uses RTS/CTS to develop collision aware systems but by increasing overheads in return. Robust Rate Adaptation Algorithm (RRAA) [4] uses frame error rate to rate change the decision. In Effective Rate Adaptation (ERA) [5] authors have proposed fragmentation to combat collisions. Some authors have proposed change in IEEE 802.11x standards Receiver Based Auto-Rate (RBAR) [6] whereas some have developed application specific solutions. Signal-to-noise ratio or received signal strength based approaches are faster adapting to dynamic channel conditions whereas transmission success or failure based approaches may take longer to adapt.

2 Metric for Channel State Information

Channel state information is a mandatory input for adaptive link measures. The four primary metrics which are considered for capturing the quality of a wireless link are:

2.1 RSSI (Received Signal Strength Indication)

According to commercial NICs available, RSSI is measured during the Physical Layer Convergence Protocol (PLCP) header and preamble which are sent at commonly supported lowest data rates. Once the header is transmitted, rest of the data received from higher layers, also called as the PLCP Service Data Unit (PSDU) is transmitted at the rate specified in the header. If PLCP header and preamble are not received due to interference, RSSI may not be recorded at all. There is also a possibility that PLCP header and preamble are not affected by interference but rest of the packet which may be sent at higher data rate is corrupted due to interference. In both cases, measurement of RSSI will not be a correct indictor of signal power received at the receiver. Experiments carried out [7] indicate that RSSI measured during simultaneous transmission from many transmitters remain stable and does not reflect the effect of interference and channel fluctuations accurately.

Figure 1a indicates that trend of RSSI may not point out interference effectively whereas (b) shows that interferer’s power may directly affect packet delivery ratio.

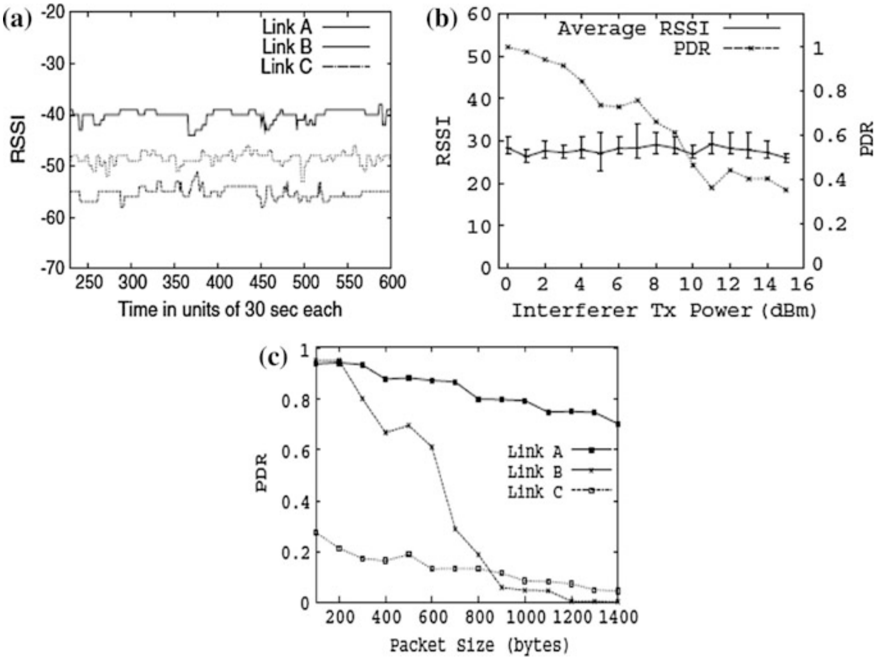


Fig. 1 **a** RSSI as interferer's power is increased on different links [7]. **b** Effect of interference on RSSI [7]. **c** Dependency of PDR packet size [7]

2.2 SINR (Signal-to-Interference-Plus-Noise Ratio)

This is one of the best parameters to evaluate the instantaneous characteristics of the link and its effect on the signal received. But SNR thresholds, low and high corresponding to 10 and 90 % frame delivery ratio respectively are, hardware dependant. This requires calibration for a pair of transmitter and receiver BERs with parameters directly proportional to channel condition which is the function of SNR. There exists a range of SNR/SINR values for which frame delivery ratio may switch from 0 to 1. But in practice all NIC cards give RSSI measurements and not SINR. Further, RSSI is measured only during preamble, that too if received, it does not reflect correct extent of interference present on the link. Trials by [7] show that PDR *(Packet-Delivery Ratio)* is directly proportional to it and is a better measure of interference than RSSI.

2.3 PDR (Packet-Delivery Ratio)

PDR is a measurable metric prevalent in accessing the link quality. But it is highly dependent on packet size. Smaller the size, higher the probability that only few bits are in error. Therefore, with heterogeneous links operating with different protocol packet sizes, PDR may not be able to represent a consistent estimate of the link quality. PDR is also dependent on the bit rate used. In lossy environment higher bit rates may have lower PDR than robust low bit rates. Figure 1c shows the same with three different links.

2.4 BER/PER (Bit/Packet Error Rate)

Analyzing the bit and packet error rate may be useful to characterize the link quality. For analysis of bit errors and packetization of errors the Markov models are widely used. BER of 10^{-5} is acceptable for wireless LAN applications. The PER values vary on different transmission parameters such as transmission power, modulation scheme (transmission rate) used, and packet size. Logically, it appears that BER will increase with higher data rates as dense constellation is prone to interference but due to fading, the BER or PER may not always monotonically decrease as the transmission rate is reduced. A practical rule for correctly estimating a BER of the order of 10^{-p} is that we need to transmit 10^{p+2} bits. This ensures approximately, an accuracy of two significant digits in the computation of BER. The BER is computed only from the received packets (correct or corrupted). Since PER is computed from received packets, packets lost due to interference and noise will not be accounted for and it may project wrong picture of link quality.

Random bit errors can be attributed to various reasons and may not provide sufficient information regarding the status of channel. But a number of either consecutive bits or packets lost or in error may indicate a poor connectivity between source and destination due to severe fading or variable length coding (VLC) techniques [8].

For a given PER single bit errors may be more damaging than a cluster of errors. This is so because a number of single errors have more number of corrupted frames leading to degradation of received video.

3 Separating Signal Attenuation from Collisions

In a wireless domain, reception of weak/no signal or packets received with errors is very frequent. This may happen due to collision of concurrent transmission or signal attenuation and fading. Corrective action is initiated to arrest further loss [9]. But two different causes require two different corrective actions, exponential back off for collisions, and change of transmission data rate or power for signal attenuation. But design of 802.11 is such that it provides only a binary feedback of success or failure for packet transmission. It implements a back off on packet loss and subsequent failures implement rate adaptation. This absence of cause detection may in fact lower the network capacity. It is suggested that metric consisting of bit, symbol error pattern, errors per symbol, and joint distributions of these could be used to separate collision losses from weak signal loss. Since this requires analysis of entire packet received in error, the packet in error has to be sent back to the sender. In [10] authors have shown experimental data analysis with interesting results.

- Figure 2 shows that if CDF of BER and SER are plotted then statistics of collision has spread over a wider range than the data from a weak signal.
- Figures 3 and 4 demonstrate that there is more number of errors per symbol in case of collision than a weak signal with larger bursts of contiguous symbols in error.

As this complex technique requires feedback to be sent to the sender, a novel method is used to make a skilled guess of loss differentiation by analyzing parameters of received signal. It is categorized as channel error if low SNR signal is received. If preamble was not received correctly then it is classified as collision

$$\text{Error Score ES} = \sum_{k=1}^{n} (\text{length of symbol error burst } k)$$

Fig. 2 CDF of BER [10]

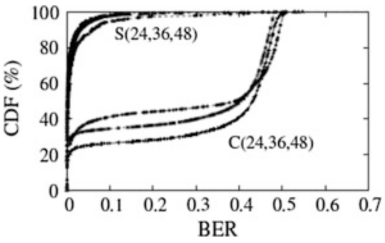


Fig. 3 CDF of error rate per symbol [10]

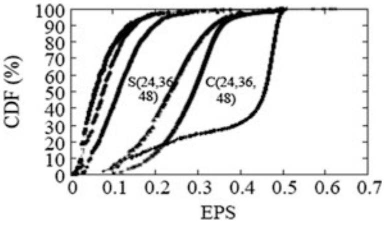


Fig. 4 CDF of Length of Symbol Error Burst for packets in error [10]

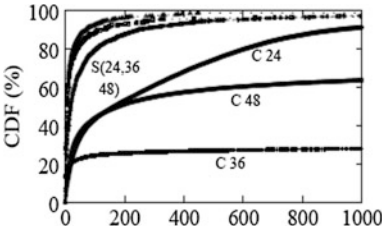
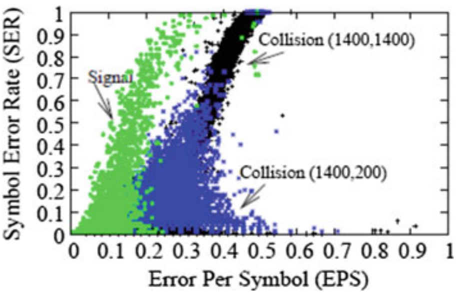


Fig. 5 Scatter plot of SER versus EPS [10]



But if preamble was received but data could not be decoded then it is an asynchronous loss. Although this procedure is straightforward there is always a possibility of false positive.

- *Joint Distribution of SER and EPS*: It is logical that packet in error due to collision has more number of errors per symbol and higher symbol error rate. Figure 5 underlines the same.

In CSMA based networks, it could be concluded [11] that

a. If interfering signal is already present when desired source starts transmission then it is taken as a class of interference for which threshold for detecting interfering signal should be reduced.
b. If interfering signal occurs after desired source starts transmission then power of signal should be increased so as to dominate and avoid corruption of source signal already started.
c. But if interfering signal starts at the same time as the desired signal then it is taken as collision and exponential back off and contention window optimized algorithm could be invoked

Another novel way that authors have tried with success is to append Cyclic Redundancy Check (CRC) after every small data segment in data packets [12]. If at the receiving end, a number of erroneous data packets go beyond a certain threshold then decision is taken in favor of collision otherwise the damage is due to channel errors. It is argued that if it was collision then chunk of sequential data would be affected and not disjointed.

Use of RTS and CTS has been used to differentiate between losses due to collision and due to channel error. But it leads to increased overheads especially in high speed wireless data transfer. But this technique can only be used for 802.11 standard and its flavors [13].

SNR and PDR are strongly correlated. There is a steady rise in PDR after threshold SNR. It becomes stable at certain SNR and saturates. This high SNR threshold can also be measured. This correlation is disturbed in presence of interference and losses could be separated from the ones due to signal attenuation.

4 Considerations for Link Rate Adaptation

To combat dynamic characteristics of wireless channel which causes packet loss, high bit error rate and delay, sending node can adapt its modulation and coding according to instantaneous channel status. But parameters required to study the channel quality, e.g., SNR fluctuates with time. It is seen in the Fig. 6a, b.

One of the techniques for link adaptation uses variable transmission or coding rate [14]. By increasing the coding rate system throughput increases but it may cause network congestion due to high bit rate. It leads to further delay and distortion. Power to be transmitted is also higher for this data rate. This calls for well designed optimum thresholds for encoding rate.

Most bit rate adaptation techniques use either frame receptions or SNR. Channel conditions that have effect rate adaptation are (i) coherence time (ii) delay spread (iii) interference, and (iv) physical layer capture.

Frame-level techniques consider percentage of lost frame and hence require several frames, to predict the channel condition. Known as loss triggered, these techniques tends to be slow especially for highly dynamic channels.

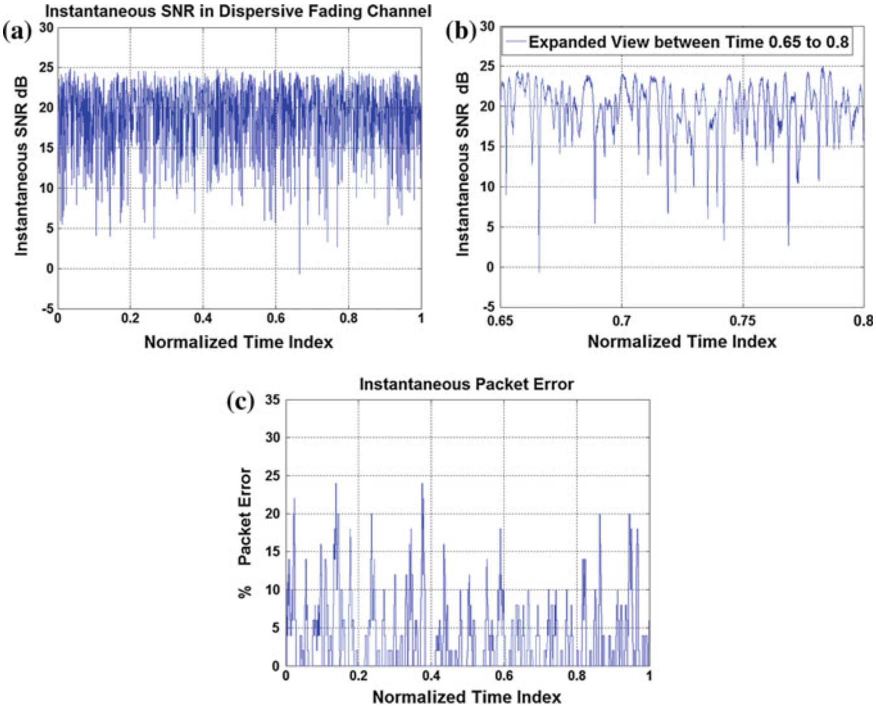


Fig. 6 **a** Continuously changing SNR in dynamic channel conditions. **b** Expanded view of SNR w.r.t. time. **c** Variable packet error rate in dynamic channel conditions

- BER and SNR are interrelated. SNR could also be used as a trigger for rate adaptation. But mobility has great impact on SNR. If it is measured at the start of the packet, reading may not be the same towards the end of the packet. Due to heterogeneous networks and dynamic channel conditions periodic training and tracking are mandatory for these protocols. It has been observed that collision losses adversely impact the performance of rate adaptation protocols. Some cross-layer wireless bit rate adaptation algorithm estimates the interference-free BER of received frames
- Figure 7 show that different frequencies undergo different fading. Heavily faded frequencies will require robust modulation, strong coding, and higher power for sufficient packet delivery ratio.

Figure 8a shows a distinct relation between SNR and PDR. Picking bit rates using SNR/BER thus estimated can be used to select appropriate bit rate. It enables to react quickly to channel variation without requiring any environment-specific calibration. BER thus estimated can be applied to a variety of wireless cross-layer protocols that, for example, allocate frequency or transmit power, or perform efficient error recovery.

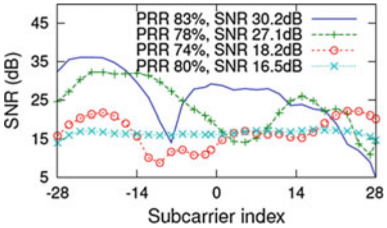


Fig. 7 Frequency selective fading for OFDM for four links with 80 % of packet delivery at 52 Mbps [15]

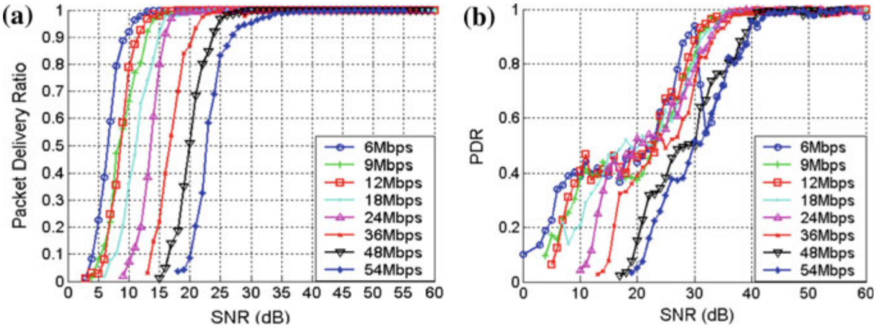


Fig. 8 **a** FDR as a function of SNR [13]. **b** Impact of Interference [13]

Different transmission rates have corresponding SNR thresholds which in turn are decided by the proprietary hardware. Higher rate requires higher SNR to sustain. There is a “transition band” in SNR where FDR goes from zero to 100 %. This transition band can be 5–7 dB wide. If interference exists in the environment, the SNR–FDR relationship may be distorted and will cause many frame losses. The transition bands are stretched and the SNR–FDR curves become irregular as seen in Fig. 8b.

SNR-based data rate adaptation based on channel state information requires Received Signal Strength (RSS) at the receiver along with estimation of noise and interference encountered. Some adaptations consider explicit feedback coming from the destination node in form of RTS/CTS but amounting to increased overhead. Authors [16] passively monitor destination, as all nodes inform other nodes in the range about the interference encountered and power they transmit. Final data rate could be decided upon removing transient fades occurring in the SINR. Adaptations, in general, quickly switch to lowest supported rates to get the packet through and extract vital information about channel state from acknowledgement received.

5 Conclusion and Future Scope

The four primary metrics which are considered for capturing the quality of a wireless link are RSSI, SINR/SNR, PDR, and BER/PER. RSSI is measured during the PLCP header and preamble sent at commonly supported lowest data rates. Tendency of RSSI may not point out interference effectively. It is one of the best parameters to evaluate the instantaneous characteristics of the link but is difficult to measure. PDR and BER are useful if analyzed correctly as they are dependent on numerous parameters and conditions.

Packet loss or corruption may be due to the collision of concurrent transmission or signal attenuation and fading. But two different causes require two different corrective actions to arrest further loss. Incorrect measure may in fact lower the network capacity.

In link adaptation bit rate selection enables to react quickly to channel variation without requiring any environment-specific calibration. It uses SNR or BER for decision making. But SNR thresholds are hardware dependant and calibration is needed for a pair of transmitters and receivers.

In link adaptation bit rate selection enables to react quickly to channel variation without requiring any environment-specific calibration. It uses SNR or BER for decision making. But SNR thresholds are hardware dependant and calibration is needed for a pair of transmitters and receivers.

Many designs suggested, need to be evaluated in real world at public sites, taking into consideration traffic pattern, user density, hidden nodes interference, etc. Mostly uplink traffic from a node to AP is considered for adaptation. But real need is at AP where massive downlink traffic, which may constitute 80 % of the total traffic, is handled. These new algorithms need to combat the link impaired due to signal fading and collisions equally well. Lastly they have to be systematically compared based on performance metrics.

References

1. R. G. Purandare, S. M. Koli, S. P. Kshirsagar, V. V. Gohokar 'Impact of Bit Error on Video Transmission over Wireless Networks and Error Resiliency' International Conference on Image Information Processing (ICIIP 2011), Nov 2011, Shimla, H.P., India, 978-1-61284-860-0/11.
2. M. Lacage, M. H. Manshaei, and T. Turletti, "IEEE 802.11 rate adaptation: a practical approach," in Proceedings of the 7th ACM Symposium on Modeling, Analysis and Simulation of Wireless and Mobile Systems (MSWiM'04), pp. 126–134, Association for Computing Machinery, October 2004.
3. S. Kim, L. Verma, S. Choi, and D. Qiao, "Collision-aware rate adaptation in multi-rate WLANs: design and implementation," Computer Networks, vol. 54, no. 17, pp. 3011–3030, 2010.
4. S. H. Y. Wong, H. Yang, S. Lu, and V. Bharghavan, "Robust rate adaptation for 802.11 wireless networks," in Proceedings of the 12thAnnual International Conference on Mobile Computing and Networking (MOBICOM '06), pp. 146–157, Los Angeles, Calif, USA, September 2006.
5. S. Biaz and S. Wu, " ERA: Effective Rate Adaptation for WLANs," in IFIP Networking 2008. Singapore: IFIP, May 2008.
6. Holland, Gavin, Nitin Vaidya, and Paramvir Bahl. "A rate-adaptive MAC protocol for multi-hop wireless networks." In Proceedings of the 7th annual international conference on Mobile computing and networking, pp. 236– 251. ACM, 2001.
7. Angelos Vlavianos, Lap Kong Law, Ioannis Broustis†, Srikanth V. Krishnamurthy, Michalis Faloutsos Invited Paper 'Assessing Link Quality in IEEE 802.11 Wireless Networks: Which is the Right Metric?' Personal, Indoor and Mobile Radio Communications, 2008. PIMRC 2008. IEEE 19th International Symposium.
8. Aderemi A. Atayero, Oleg I. Sheluhin and Yury A. Ivanov 'Modeling, Simulation and Analysis of Video Streaming Errors in Wireless Wideband Access Networks'IAENG Transactions on Engineering Technologies, DOI:10.1007/978-94-007-4786-9_2, Springer Science + Business Media Dordrecht 2013.
9. Lito Kriara, Sofia Pediaditaki, Mahesh K. Marina 'On the Importance of Loss Differentiation for Link Adaptation in Wireless LANs' 2nd International Conference on Broadband Networks, 2005. BroadNets 2005. Oct. 2005, Page(s):659–667 Vol. 1.
10. S. Rayanchu, A. Mishra, D. Agrawal, S. Saha, S. Banerjee,' Diagnosing Wireless Packet Losses in 802.11: Separating Collision from Weak Signal' In Proc. IEEE INFOCOM, Phoenix, AZ, April 2008.
11. H. Ma, J. Zhu and S. Roy, On 'Loss Differentiation for CSMA-Based Dense Wireless Network' In IEEE Communications Letters, Nov 2007. Pages 877–879, DOI:10.1109/LCOMM.2007.071154.
12. G. Kyriakou et.al. 'A Framework and Experimental Study for Discrimination of Collision and Channel Errors in Wireless LANs' TridentCom 2011, Shanghai, China, Page(s):271–285.
13. J Zhang, K Tan, J Zhao, H Wu, Y Zhang,' A Practical SNR-Guided Rate Adaptation'INFOCOM 2008. 978-1-4244- 2026-1/08.
14. Ali Abdulqader Bin-Salem and Tat Chee Wan 'Survey of Cross layer Designs for Video Transmission Over Wireless Networks' IETE TECHNICAL REVIEW Vol 29| Issue 3, May 2012.

15. D. Halperin, W. Hu, A. Shethy, D. Wetherall “Predictable 802.11 Packet Delivery fromWireless Channel Measurements” SIGCOMM’10, New Delhi, India. Copyright 2010 ACM 978-1-4503-0201-2/10/08.
16. G. Judd, X. Wang and P. Steenkiste, ‘Efficient Channel–Aware Rate Adaptation in Dynamic Environments’, In Proc. ACM MobiSys, Breckenridge, Colorado, June 2008.

N-Gram Classifier System to Filter Spam Messages from OSN User Wall

Sneha R. Harsule and Mininath K. Nighot

Abstract Online Social Network (OSN) allows users to create, comment, post, and read articles of their own interest within virtual communities. They may allow forming mini-networks within the bigger, more diverse social network service. But still, improper access management of the shared contents on the network may give rise to security and privacy problems like spam messages being generated on someone's public or private wall through people like friends, unknown persons, and friends of friends. This may also reduce the interest of Internet data surfing and may cause damage to less secure data. To avoid this, there was a need of a system that could remove such unwanted contents, particularly the messages from OSN. Here in this paper, for secure message delivery I have presented a classifier system based on N-gram generated profile. This system consists of ML technique using soft classifier, that is, N-gram which will automatically label the received messages from users in support of content-based filtering. Effectiveness of N-grams is studied in this paper for the purpose of measuring the similarity between test documents and trained classified documents. As an enhancement, N-gram method can also be used to detect and prevent leakage of very sensitive data by using N-grams frequency for document classification.

Keywords OSN · Access management · Unwanted contents · N-gram · Soft classifier · Machine learning technique, etc.

S.R. Harsule (✉) · M.K. Nighot
Department of Computer Engineering, KJ College of Engineering and Management Research, Savitribai Phule Pune University, Pune, Maharashtra, India
e-mail: hsneha30@gmail.com

M.K. Nighot
e-mail: imaheshnighot@gmail.com

H.S. Saini et al. (eds.), *Innovations in Computer Science and Engineering*,
Advances in Intelligent Systems and Computing 413,
DOI 10.1007/978-981-10-0419-3_3

1 Introduction

The most popular and interactive medium for the purpose of sharing information is OSN. Social networks have become the hottest online trend in the past few years. People around the world are now able to communicate and share information easily among themselves. Social networks formed by users provide a platform for finding and managing useful information. Contents like text, image, audio, and video data are now possibly exchanged. The Users Personal data in OSN requires high level of privacy and security. To preserve privacy for OSN data there is a need to improve access control over the contents displayed on the web pages. There are different Access Control Models (ACM) among which some are OSN-specific ACMs [1]. As OSN users still lack ACM functionality to avoid spam messages posted on their wall, classifier system introduced in this paper is the best solution to this.

Contents exchanged over network are mostly in the form of text. Text mining is a technique made in association with information retrieval, machine learning and statistics, and data mining [2]. Text mining has gained high commercial and potential value. Information filtering and Information retrieval are two sides of the same coin [3]. Text mining can also be referred as text data analysis or data mining in which important information is possible to be retrieved from text. To do text preprocessing or prerequisites, the phases are parsing, tokenization, stemming, etc., which I am going to use in my project. Method proposed in this paper starts extracting N-grams of variable length from OSN messages which further generate N-gram profiles. N-gram profiles are generated for both test dataset and trained dataset [4]. Then by using any of the similarity measures like Jaccard, cosine, Dice, etc., we can compare test and trained data. Based on this distance measure, the classification of document will be made. Exploitation of Machine Learning (ML) text categorization techniques will be made for automatically assigning a set of categories based on contents of each text message. The classification will be spam and unspam OSN messages. Thus, the aim of this paper is divided into three parts: first, to study the effectiveness of using N-gram based system for classification of messages. Second, to define an optimum N-gram size along with an optimum category profile size to achieve greater accuracy, and finally, to examine the effect of modification in the document over the system. In addition, I am possibly trying to give a short text message classification support to my proposed system and also a black list mechanism in which a user can decide on who can post messages on their walls. In short, my work is to present and experimentally evaluate an automated system called as Filtered Wall (FW).

2 Related Work

To provide secure access to user's data, this paper focuses on filtering unwanted OSN messages; so it was beneficial to study existing techniques. Here, literature review work of some previous papers has been performed related to my topic. With the support of this study, we have tried to build classifier system which is more effective to some extent.

Jin et al. introduced a spam detection system for social network based on data mining in [5]. The advantage of using social network is to spread information across globe quickly and effectively. Besides this advantage, some issues also discussed in this paper are unrelated information generation, security and privacy issues in commercial use of social network, etc. They have proposed scalable spam detection system for OSN security. To achieve this they used GAD clustering algorithm in support of active learning algorithm to deal with large-scale data clustering and real-time detection challenges. They designed a framework, in which they first collected historical social media data and extracted both the content and the social network features, and performed active learning algorithm to build classification model and found spams. In the next stage, they monitored real-time activity of the social network and performed an online active learning algorithm; on that basis they made predictions and sent alarm notification to the clients about spam message detection. They also tried to collect feedback from clients and updated the model according to it. But such classification system is not working up to the mark for all OSN websites because sites like Facebook contain large amount of spam message datasets which are relatively difficult to handle. Alneyadi et al. performed word N-gram-based classification in order to avoid data leakage [6]. They proposed this work to reduce the problem of data leakage, because it may harm an individual or an organization. They tested 180 articles among which 142 were classified correctly. They designed the DLP (Data Leakage Prevention) system but lack of functionality in automatic category profile update may be considered as the future work for which ML algorithm will be required. Vanetti et al. designed a system to filter unwanted messages from OSN user wall [7]. They designed three-layer framework in which top layer was Graphical User Interface (GUI) for the OSN user, middle layer was processing system divided as content-based message filtering module and short text message classification module and the bottom layer comprised OSN database, social graph, user profile handled by social manager, etc. They enforced to classify short text messages to design strong filtering system. In support of the content-based filtering they also provided black lists (BL) mechanism and filtering rules (FR) flexibility to the user. Before directly going into the proposed classifier system study, preference was given to go through various sources of information that come under the domain called data mining. Spam messages are identified and also their diverse effects that can harm any individual or an organization are read. So basically, spam is a subset of electronic spam which may consist of identical messages sent to a number of users by clicking on the links in unwanted emails. By clearing such prerequisites concepts and reading history of spam messages mentioned in some articles [8],[9]. I came across some interesting incidences that inspired me to focus and work on data mining domain and to design a classifier system that will help to reduce diverse effects of spam to some extent. Since the early 1990s, email spams have gradually increased with the help of a network of computers affected by virus called Botnets. Hence I felt that, spam messages should be filtered and separated from the original messages or legal messages.

3 System Design

To know more in detail about the proposed work, it gets divided here into three parts: Architecture, the Algorithm used, and the mathematical representation of the system.

3.1 Workflow of Architecture

As shown in Fig. 1 our system is divided into three main modules which are Social Management, N-gram classifier, and GUI. All these modules are explained in detail below:

Social Management It will handle database-related activities like storing trained dataset, Fetching API from OSN, storing messages, data classification results, etc. **N-Gram Classifier** Some preprocessing activities such as labeling, categorization, parsing, and stemming on text will be performed in this system part. Classification

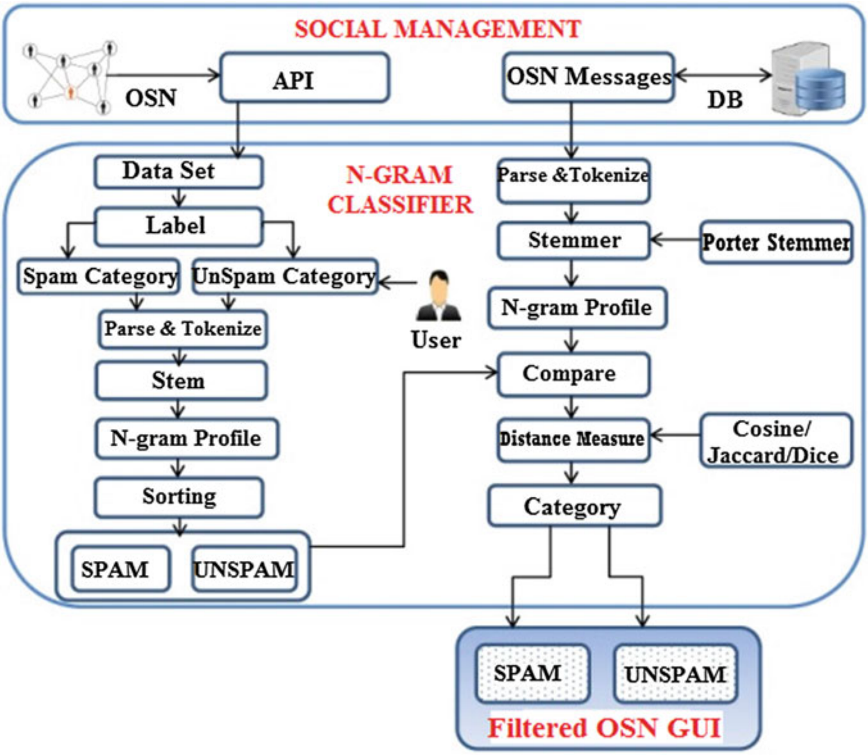


Fig. 1 Architecture of N-Gram classifier system

of spam and unspam messages will be based on the comparison. The system will generate N-gram profiles of test and trained message sets using similarity techniques such as Jaccard [10]. Proposed system has used Jaccard to calculate the distance measure.

Graphical User Interface (GUI) It will be the communication interface between user and system where user will get filtered OSN wall due to the classifier used here.

3.2 Algorithms

Proposed system has used two algorithms. First algorithm is used to find root words from message which is of well-known standard, Porter Stemmer Algorithm [11]. Our main task in this paper was performed by the next algorithm called N-Gram Classification, in which OSN messages get classified as spam and unspam messages.

N-Gram classification algorithm

```
Step-1: Fetch OSN API to obtain training dataset.
Step-2: Select appropriate value of N-Grams.
Step-3: Stop words must be removed from message.
Step-4: Apply stemming operation to these messages.
Step-5: Now append Strings of stem words.
Step-6: For i=0 to total words in message.
                    For j= i+1 to N Create N-Gram String S.
                              Add String to the vector V.
                    End (until to N)
               End For loop.
Step-7:   Display N-Grams profiles of trained messages.
Step-8:   Sort out these profiles as Spam and Unspam as our final predefined
          trained Data sets.
Step-9:   Repeat step-2 to step-7 for OSN messages means for test dataset.
Step-10: Now Compare N-Gram generated profiles of  trained, test messages.
Step-11: Calculate distance measure using Similarity techniques.
Step-12: With smallest distance measure content classified as spam and Unspam.
Step-13: End
```

4 Results

The design of the system is in terms of precision, recall, and F-measure represented below: Let us assume UP = unspam positive messages; SP = spam positive messages; UN = unspam negative messages and SN = spam negative messages, then expected result of system is shown in Tables 1 and 2. Table 1 contains actual

Table 1 Precision and recall

Dataset	Actual messages		Correctly retrieved		Incorrectly retrieved	
			N-Gram	TF/IDF	N-Gram	TF/IDF
UNSPAM	10 =	5	5	4	0	1
SPAM		5	4	4	1	1
UNSPAM	15 =	10	9	8	1	2
SPAM		5	4	3	1	2
UNSPAM	20 =	10	9	7	1	3
SPAM		10	8	8	2	2
UNSPAM	25 =	15	12	12	3	3
SPAM		10	10	8	0	2
UNSPAM	30 =	15	14	13	1	2
SPAM		15	13	12	2	3

messages, correctly retrieved messages, and incorrectly retrieved messages are calculated.

Table 2 shows expected accuracy result of the proposed system in comparison to the existing one [12].

To know the accuracy of the system, Fig. 2 is designed based on the values shown in the tables.

Hence, it becomes easy to predict the accuracy of the system proposed in this paper.

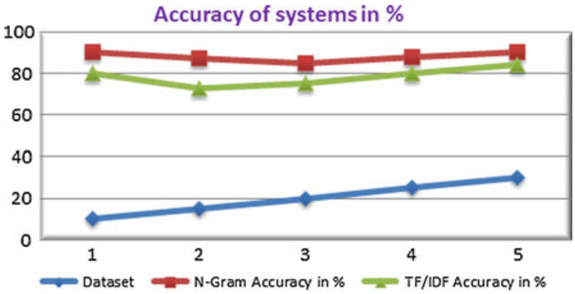


Fig. 2 Graph for accuracy

Table 2 Predicted accuracy result

Dataset (spam + non spam)	N-Gram accuracy in %	TF/IDF accuracy in %
10	90	80
15	87	73
20	85	75
25	88	80
30	90	84

5 Conclusion and Future Work

Here, we have presented N-gram classifier system which helps to filter spam and unspam messages effectively. The approach we have taken to implement the proposed work is first, collecting and categorizing trained documents as spam and unspam text categories based on ML process. We did some text preprocessing activities on it and received test dataset from OSN. We then created the N-Gram Profiles for both of them. Similarity technique-based comparison was made which classified the document as spam or unspam messages based on its distance measured. Such classification will definitely be advantageous for OSN users to have reliable access to their contents. In the future, there is a lot of scope to make advancement in the system by considering many of the aspects such as black-list users and short text classification mechanism for making the system automated and powerful. Also, there will be a possibility to design advanced classifier system which may consist of a unique and useful combination of different data mining features, algorithms, learning, ACM models, etc.

References

1. Rula Sayaf and Dave Clarke, “Access Control Models For Online Social Networks”, in book, internationally recognised scientific publisher, IGI Global, [2012].
2. Mrs. Sayantani Ghosh, Mr. Sudipta Roy, and Prof. Samir K. Bandyopadhyay, “A tutorial review on Text Mining Algorithms”, in International Journal of Advanced Research in Computer and Communication Engineering, Vol. 1, Issue 4, June [2012].
3. Nicholas J. Belkin and W. Bruce Croft, “Information filtering and information retrieval: Two sides of the same coin?”, in Communications of the ACM v35 n12p29(10), Dec [1992].
4. Zakaria Elberrichi & Badr Aljohar, “N-grams in Texts Categorization”, in Scientific Journal of King Faisal University Vol. 8 [2007].
5. Xin Jin, Cindy Xide Lin, Jiebo Luo and Jiawei Han, “A Data Mining based Spam Detection System for Social Media Networks”, in Proceedings of the VLDB Endowment, Vol. 4, No. 12, August 29th - September 3rd [2011].
6. Sultan Alneyadi, Elankayer Sithirasenan and Vallipuram Muthukkumarasamy, “Word N-gram Based Classification for Data Leakage Prevention”, in 12th IEEE interanational conference July [2013].
7. Marco Vanetti, Elisabetta Binaghi, Elena Ferrari, Barbara Carminati, and Moreno Carullo, “A System to Filter Unwanted Messages from OSN UserWalls”, in IEEE Transactions On Knowledge And Data Engineering, Vol. 25, No. 2, February [2013].
8. Salwa Adriana Saab, Nicholas Mitri and Mariette Awad, “Ham or Spam? A comparative study for some Content-based Classification Algorithms for Email Filtering”, in 17th IEEE Mediterraneaan Electronical Conference,Beirut,April [2014].

9. Christina V, Karpagavalli S and Suganya G, "A Study on Email Spam Filtering Techniques", International Journal of Computer Applications Vol. 12– No.1, December [2010].
10. RoissAlhutaish and Nazlia Omar, "Arabic Text Classification Using K-Nearest Neighbour Algorithm", in The International Arab Journal of Information,vol.12,No.2,March [2015].
11. Presentation,Porter Stemmer Daniel Waegel CISC889/ Fall [2011].
12. Taher Zaki, Youssef Es-saady, Driss Mammass, Abdellatif Ennaji and Stéphane Nicolas, "A Hybrid Method N-Grams-TFIDF with radial basis for indexing and classification of Arabic documents", in International Journal of Software Engineering and Its ApplicationsVol.8, No.2, pp.127-144,[2014].

A Recent Study of Emerging Tools and Technologies Boosting Big Data Analytics

Govind Pole and Pradeepini Gera

Abstract Traditional technologies and data processing applications are inadequate for big data processing. Big Data concern very large-volume, complex formats, growing data sets with multiple, heterogeneous sources, and formats. With the reckless expansion in networking, communication, storage, and data collection capability, the big data science is rapidly growing in every engineering and science domain. Challenges in front of data scientists include different tasks, such as data capture, classification, storage, sharing, transfer, analysis, search, visualization, and decision making. This paper is aimed to discuss the need of big data analytics, journey of raw data to meaningful decision, and the different tools and technologies emerged to process the big data at different levels, to derive meaningful decisions out of it.

Keywords Big data · Data mining · Clustering · Data volume · Data velocity · Data variety · Big data mining · Big data analytics

1 Introduction

The term, Big data is coined to define a huge volume of organized and unorganized data in a multiple format, that is outsized to process using classical technologies, algorithms, databases, and processing methods [1, 2]. The Evolving technology platforms provide the competency to process data of various formats and structures without worrying about the boundaries related to traditional systems. Data collected from different processes is used to make decisions feasible to the understanding and

G. Pole (✉) · P. Gera
Department of Computer Science and Engineering, K L University,
Guntur, A.P., India
e-mail: govindpole@gmail.com

P. Gera
e-mail: pradeepini.gera@gmail.com

H.S. Saini et al. (eds.), *Innovations in Computer Science and Engineering*,
Advances in Intelligent Systems and Computing 413,
DOI 10.1007/978-981-10-0419-3_4

requirements of those executing and consuming the results of the process [1]. This administrative behavior was the underlying theme for Herbert Simon's view of bounded rationality. The dispute in decision making and administrative behaviors makes a sense, as we bound the data in the process of modeling, applying algorithmic applications, and have always been seeking discrete relationships within the data as opposed to the whole picture [3]. The big data analytics evolved to support the decision making process by collecting, organizing, storing, retrieving, and managing big data.

1.1 Need for Big Data Analytics

Globalization and personalization of services are some crucial changes that motivate Big Data Analytics. Many issues in real applications need to be described using a graphical structure such as the optimization of railway paths, prediction of disease outbreaks, and emerging applications such as analysis of social network, semantic network, analysis of biological information networks, etc. [4]. Figure 1 shows the hierarchy of data analysis work.

1.2 Big Data Processing

Conceptually there are two ways of data processing that are accepted as a de facto standard. In centralized processing collected data is stored at a centralized location and then processed on a single machine/server. This centralized machine/server has very high configurations in terms of memory, processor, and storage. This

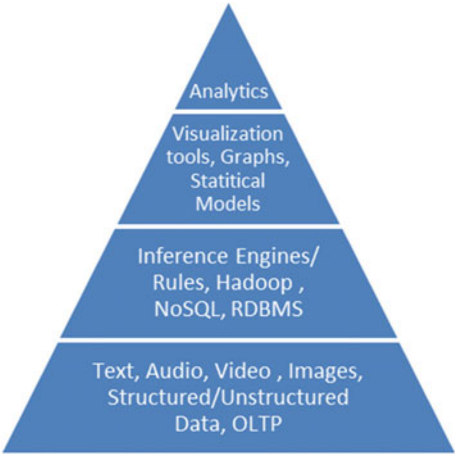

Analytics
Visualization tools, Graphs, Statitical Models
Inference Engines/ Rules, Hadoop , NoSQL, RDBMS
Text, Audio, Video , Images, Structured/Unstructured Data, OLTP

Fig. 1 Data layers in big data analytics

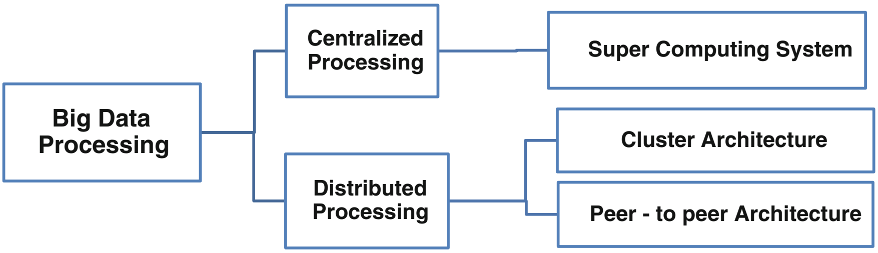


Fig. 2 Fundamental styles for data processing

architecture is well suited for small organizations with one location for production and service, but for large organizations having different geographical locations is almost out dated. Distributed processing evolved to overcome the limitations of the centralized processing, where there is no need to collect and process all data at a central location. There are several architectures of distributed processing. For example Cluster architecture and Peer-to-peer architecture (Fig. 2).

2 Big Data Analysis Steps

Big data involves data sets whose size is challenging to present tools and technologies to manage and process the data within acceptable time [5]. Journey of raw data to make useful decisions using big data analytics is governed by the following steps (Fig. 3).

2.1 Data Collection

Data collection is a process of retrieving raw data from real-world data sources. The data are fetched from different sources and stored to a file system. Inaccurate data collection can affect the succeeding data analysis procedure and may lead to unacceptable results, so the data collection process should to be carefully designed [6]. In the modern world, the distributed data collection from different geographical systems and networks is pervasive and there is a demand to discover the hidden patterns of the data stored with these heterogeneous or homogeneous distributed nodes [7]. Useful Tools: Chukwa, WebCrawler, Pig, and Flume.

2.2 Data Partition

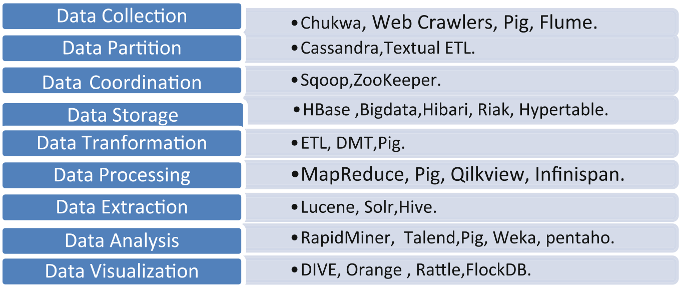
To handle very large data volume different partitioning techniques like data tagging, classification, and incremental clustering approaches are used. Many clustering approaches have been designed to help address a large data [7, 8]. Useful Tools: Textual ETL, Cassandra. To minimize the complexity of processing large data several algorithms are emerged. These algorithms are categorized as Text mining, Data mining, Pattern processing, Mahout, Scalable nearest Neighbor Algorithms [9].

2.3 Data Coordination

Coordination governs the exchange of data among the data warehouses, relational databases, NoSQL Databases, and different big data technologies. For example, Sqoop [10] can be used to exchange data to and from relational databases like MySQL or Oracle to the HDFS. The cross-cloud service requires data aggregation and transfer among different systems to process large-scale big data from different sources [11]. Flume is a distributed system that manages large amount of data from different systems and networks by efficiently collecting, aggregating, and moving it to a desired location [12]. On the other hand the Zookeeper is a centralized service which provides distributed synchronization, naming, and group services. It also maintains the configuration information of the systems. The APIs are available for various programming languages like Java, Perl, and Python [13].

2.4 Data Transformation

Data transformation refers to the conversion of data values in one format of source system to another format of destination system. Useful tools: ETL, DMT, and Pig.



Data Collection	• Chukwa, Web Crawlers, Pig, Flume.
Data Partition	• Cassandra,Textual ETL.
Data Coordination	• Sqoop,ZooKeeper.
Data Storage	• HBase ,Bigdata,Hibari, Riak, Hypertable.
Data Tranformation	• ETL, DMT,Pig.
Data Processing	• MapReduce, Pig, Qilkview, Infinispan.
Data Extraction	• Lucene, Solr,Hive.
Data Analysis	• RapidMiner, Talend,Pig, Weka, pentaho.
Data Visualization	• DIVE, Orange , Rattle,FlockDB.

Fig. 3 Different steps in big data analysis

DMT is a Data Migration Tool from TCS. It provides accurate data migration from conventional databases to Hadoop repository like HDFS or Hbase.

2.5 Data Storage

The main challenge in front of scalable bigdata system is the need to store and manage large volume of data collected. At the same time, it should provide methods and functions for efficient access, retrieval, and manipulation to diverse datasets. Due to a variety of dissimilar data sources and large volume, it is tough to gather and integrate data with performance guarantee from distributed locations. The big data analytic system must effectively mine such huge datasets at different levels in real time or near real time [6]. Useful Tools: Hbase [14], NoSQL [15], Gluster [16], HDFS [17], GFS [18].

2.6 Data Processing

There is no fixed data processing model to handle the big data because of its complex nature. The data processing in big data analytics is rather schema less or non-structured. To process the structured and unstructured data in different format, a mixture of various technologies like. Hadoop, NoSQL, and MapReduce is used. MapReduce framework is popular for processing and transformation of very large data sets [5]. The Qlikview is an example of In-memory data processing solutions for big data which is very useful for advanced reporting. Infinispan is an extremely scalable and highly available data processing grid platform [19].

2.7 Extract Data

This phase extracts data from the files and databases and produces result set, which is used for further analytics, reporting, integration, and visualization of the result.

Data Query Tool Hive is a data query tool used for ad hoc queries, data summarizations, data extraction, and other analysis of data. It uses HiveQL language which supports the flexible querying to the big data [20].

Big Data Search In order to scale to huge data sets, we use the approach of distributing the data to multiple machines in a compute cluster and perform the nearest neighbor search using all the machines in parallel [9]. Lucene [21], Solr [22] are the major examples of such search tools.

2.8 Data Analysis

RapidMiner is an open source system that uses data and text mining for predictive analysis [23]. Pentaho is an open source business intelligence product which provides data integration, OLAP services, reporting, and ETL capabilities. Talend [24] and SpagoBI are examples of widely used open source BI tool [25]. WEKA tool, a part of large machine learning project, consists of data mining algorithms that can be applied for data analysis. The machine learning and data mining communities have developed different tools and algorithms for tackling all sorts of analysis problems [26].

2.9 Data Visualization

The Big data is stored in files and not in table structure or format, so it is less interactive in a visualization situation. Because of this, big data need statistical softwares like R, SAS, and KXEN, where the predefined models for different statistical functions can be used for data extraction and results are integrated for statistical visualizations. Some of the well-known visualization tools are DIVE [27], and Orange [28].

3 Programming for Big Data

There are various programes useful for big data processing. The most prominent and powerful languages are R, Python, Java, Storm, etc. Python is very useful to programers for doing statistics and efficient manipulation of statistical data. This includes typically vectors, lists, and data frames that represent datasets organized in rows and columns, Whereas R outshines in the range of statistical libraries it compromises. Almost all statistical tests/methods are part of an R library. It is very easy to learn language with a vast amount of inbuilt methods. NumPy library in Python encompasses homogeneous, multidimensional array that offers various methods for data manipulation. It has several functions for performing advanced arithmetic and statistics [29]. Java and Storm [30] also play a major role in big data programming.

4 Conclusion

A changed paradigm of modern business and an advancement in communication and technology has given a new face to the analytics. Like the light fastening in computers, now people need superfast decision making and it is possible with big data analytics. Advancement in tools and technologies made it possible. Best practices have emerged to help big data processing. An interesting fact is that many of these practices are the new empowered, flexible extensions of the old one. The main thing behind the popularity of big data analysis is that it helps an organization to take corrective actions and useful decisions without much of data processing latency. Thus, big data enables decision capacity, nearly in run time environment.

References

1. X. Wu, X. Zhu, G.-Q. Wu, and W. Ding, “Data mining with big data,” *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 1, pp. 97–107, 2014.
2. L. Wang, K. Lu, P. Liu, R. Ranjan, and L. Chen, “IK-SVD: Dictionary Learning for Spatial Big Data via Incremental Atom Update,” vol. XX, no. Xx, pp. 1–12, 2014.
3. M. Augier, “Sublime Simon: The consistent vision of economic psychology’s Nobel laureate,” *J. Econ. Psychol.*, vol. 22, no. 3, pp. 307–334, 2001.
4. Y. Liu, B. Wu, H. Wang, and P. Ma, “BPGM : A Big Graph Mining Tool,” vol. 19, no. 1, 2014.
5. S. Meng, W. Dou, X. Zhang, J. Chen, and S. Member, “KASR : A Keyword-Aware Service Recommendation Method on MapReduce for Big Data Applications,” vol. 25, no. 12, pp. 1–11, 2013.
6. D. Takaishi, S. Member, H. Nishiyama, and S. Member, “in Densely Distributed Sensor Networks,” vol. 2, no. 3, 2014.
7. P. Shen and C. Li, “Distributed Information Theoretic Clustering,” vol. 62, no. 13, pp. 3442–3453, 2014.
8. Y. Wang, L. Chen, S. Member, and J. Mei, “Incremental Fuzzy Clustering With Multiple Medoids for Large Data,” vol. 22, no. 6, pp. 1557–1568, 2014.
9. M. Muja, “Scalable Nearest Neighbour Methods for High Dimensional Data,” vol. 36, no. April, pp. 2227–2240, 2013.
10. (2012), sqoop [online]. Available: https://sqoop.apache.org/docs/1.4.2/SqoopUserGuide.htm.
11. W. Dou, X. Zhang, J. Liu, J. Chen, and S. Member, “HireSome -II : Towards Privacy-Aware Cross- Cloud Service Composition for Big Data Applications,” vol. 26, no. 2, pp. 1–11, 2013.
12. (2013), Flume [online]. Available: https://flume.apache.org/FlumeUserGuide.html.
13. (2014), Zookeeper [online]. Available: https://zookeeper.apache.org/releases.html.
14. (2013). *HBase* [Online]. Available: http://hbase.apache.org/.
15. R. Cattell, “Scalable SQL and NoSQL data stores,” *SIGMOD Rec.*, vol. 39, no. 4, pp. 12_27, 2011.
16. (2014), Gluster [online]. Available: http://www.gluster.org/.
17. (2013). *Hadoop Distributed File System* [Online]. Available: http://hadoop.apache.org/docs/r1.0.4/hdfsdesign.html.
18. S. Ghemawat, H. Gobioff, and S.-T. Leung, “The Google file system,” in *Proc. 19th ACM Symp. Operating Syst. Principles*, 2003, pp.29_43.
19. (2015), Infinispan [online]. Available: http://infinispan.org/documentation/.

20. A. Thusoo *et al.*, “Hive: A warehousing solution over a Map-Reduceframework,” *Proc. VLDB Endowment*, vol. 2, no. 2, pp. 1626_1629, 2009.
21. (2014), Lucene [online]. Available: https://lucene.apache.org/.
22. (2013). Solr [Online]. Available: http://lucene.apache.org/solr/.
23. (2013). *Rapidminer* [Online]. Available: https://rapidminer.com.
24. (2015). Talend [Online]. Available: https://www.talend.com/.
25. (2015). SpagoBI [Online]. Available: http://www.spagobi.org/.
26. D. Breuker, “Towards Model-Driven Engineering for Big Data Analytics – An Exploratory Analysis of Domain-Specific Languages for Machine Learning,” *2014 47th Hawaii Int. Conf. Syst. Sci.*, pp. 758–767, 2014.
27. S. J. Rysavy, D. Bromley, and V. Daggett, “DIVE: A graph-based visual-analytics framework for big data,” *IEEE Comput. Graph. Appl.*, vol. 34, no. 2, pp. 26–37, 2014.
28. (2015). Orange [Online]. Available: http://orange.biolab.si/.
29. P. Louridas and C. Ebert, “Embedded analytics and statistics for big data,” *IEEE Softw.*, vol. 30, no. 6, pp. 33–39, 2013.
30. (2015). *Storm* [Online]. Available: http://storm-project.net/.

Registration Plate Recognition Using Dynamic Image Processing and Genetic Algorithm

Glence Joseph and Tripty Singh

Abstract Registration plate recognition plays a vital role in numerous applications in today's world. We also introduce a new approach using genetic algorithm to figure out the registration plate location. Fluctuating illumination conditions are taken care-of by adaptive threshold method. Connected element tagging is used to identify the objects in blindfolded regions. A matrix of invariant scale geometry is used for better system adaptability when applied to different plates. The convergence of genetic algorithm is greatly improved by the introduction of a newly created mutation and crossover operators. We also modify genetic algorithm to overcome the drawbacks of connected element method by importing partial matching of the characters. Finally, we take a look at the real-time challenges and remedies to it.

Keywords Image processing · Genetic algorithm · Adaptive threshold · Connected element tagging · Mutation · Crossover · Convergence · Matrix · Artificial intelligence

1 Introduction

Registration plate recognition in real-time plays a significant role in many real-world applications. These systems are employed in areas such as parking area, restricted areas for security, road traffic monitoring, bottleneck management, and toll management automation [1]. It is a challenging problem, due to the diversity in formats of the plate, different scales, angle of rotations, and nonuniformity in illumination conditions during image acquisition. Registration plates usually

G. Joseph (✉) · T. Singh
Amrita School of Engineering, Bangalore, Karnataka, India
e-mail: glencecr7@gmail.com

T. Singh
e-mail: tripty_smart@yahoo.co.in

H.S. Saini et al. (eds.), *Innovations in Computer Science and Engineering*,
Advances in Intelligent Systems and Computing 413,
DOI 10.1007/978-981-10-0419-3_5

contain multiple colors. different languages, and also different fonts. Some plates may have one color in the background while others may have background images. Registration plate locating section and a registration plate number identification section are two parts of the defined problem. Registration plate recognition algorithm consists of segmentation, extraction, and recognition divisions [2]. Character recognition system uses morphological operations, histogram manipulation along with edge detection procedures for plate localization, and characters segmentation [3]. Our target is to find a system that is more accurate on recognized registration plate.

2 Proposed Method

The proposed method introduces dynamic image and video processing methods to meet the challenges. The captured frames from the video camera are subjected to image processing techniques to localize and extract the registration plate from it. The proposed method has a futuristic genetic algorithm to locate the registration plate in the captured image. Genetic algorithm proposed has specially created mutation and crossover operators for the problem adaptability. The method made use of MATLAB to give an adaptive process for the thresholding to meet the real-time challenges. Connected element tagging was used to separate the objects in the located plate region. Universal boxing element matrix was used for adaptability of the algorithm to variety of plates. Mutation and crossover operators introduced are one of a kind, which gives the system faster convergence with minimal number of generations. Partial match of the plates with confidence percentage values was an additional feature to this proposal. Proposed system takes a look at the real-time challenges and remedies to it.

3 Implementation

The captured image is a color image keeping an account of the factor that other relevant information about the vehicles are known. RGB image has three channels. A single-point color representation of a coordinate system and a subspace in that system gives an idea of a grayscale image. In YIQ color space, the luminance (Y) represents the grayscale information, (I) represents hue and saturation, and (Q) represents the color information (Fig. 1).

$$0.298 * \mathrm{R} + 0.587 * \mathrm{G} + 0.114 * \mathrm{B}$$

Each pixel of a binary image has a maximum of two possible values, which is now a digital resembling image. Grayscale image obtained in the previous stage is converted to binary image. This is one of the critical stages in the process because

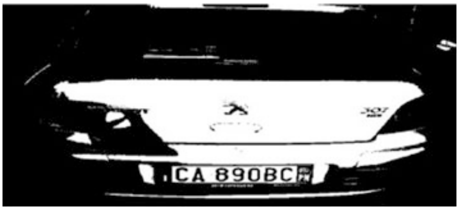


Fig. 1 Converted grayscale image in MATLAB

of variations coming across based on temporal and spatial factors in the surrounding and plate [4]. Binary conversion of image based on threshold may not see through all the challenges. Morphological process consists of erosion and dilation of the binary image obtained in the previous step. Dilation and erosion help to eliminate the holes and openings in the image. It means to say that zeroes in the binary image are mostly of irrelevant region and can be eliminated. This helps in narrowing down the search space. Connected elements tagging groups pixels of an image into elements based on the connectivity of the pixels. Pixels are tagged with a gray level or any color according to the element it was assigned to [5]. Extracting and tagging of disjoint and connected elements in an image is a key factor in many automated image analysis applications. Every object that is within the binary image is found out using an eight-point connected element analysis. N objects array is the output from stage. Objects extracted from the connected element analysis stage are strained on the basis of their widths, Widthobj, and heights, Heightobj, for the reason that the dimensions of the registration plate symbols lie between thresholds as follows: Widthmin ≤ Wifthobj ≤ Widthmax and Heightmin ≤ Heightobj ≤ Heightmax.

4 Designing of Genetic Algorithm Phase

A license plate has many complex objects inside it. It has to be mapped based on it as well. The complex variables we use are heights and widths of the rectangle bounding box as well as the top corners of the complex objects in the license plate. Fitness function is defined as the inverse of the objective distance between the prototype chromosome and the current chromosome. Random values are used inside gene. Two types of geometrical relations are used that can be defined as position correlation, represented by the relative distances between the hopping boxes of the two objects in the X and Y directions and the size correlation, represented as the relative differences in their hopping boxes between heights and widths. The objective distance functions in this genetic algorithm problem could be minimized. Universal objective distance function ObjDistk, p which is used to represent the distance between any chromosome k and the prototype chromosome p. The stochastic universal sampling method is implemented for the selection of offspring in the new generation. In this process each individual is mapped to continuous segments of line which is equal in size to its fitness as in roulette-wheel selection. A number of similarly spread out pointers are placed over the line depending on the ratio of individuals to be selected. Substitution mutation operator in which random position in the chromosome is nominated and the corresponding allele is changed by a new random object from the M available objects. Swap mutation operator does reciprocal exchange mutation which selects a couple of genes randomly and swaps them. Single-point crossover, three-parent crossover, two-point crossover, n-point crossover, uniform crossover, and alternate crossover are different types of crossover operators. These operators generate repeated genes which is not suitable for our problem. We created a new crossover operator for enhancing the chromosomes generated. In the new solution we divide the recombined chromosomes to their axis positions into two groups and sort them on the basis of the x-positions. Stopping conditions are given to genetic algorithm that stops the process as we assume that optimized fitness function is attained by this time. The two conditions we applied were first, the objective distance of the best chromosome is below six and second, if the average objective distance is not improved for successive six generations.

5 Results and Analysis

Experiments were carried out for various cameras to object relative positions in different lighting conditions. Scaling will not affect the results if done on the same dataset as long as the candidate symbols lie inside the definite ranges in the size straining stage. Sample images reveal the robustness and distinction of the proposed approach (Figs. 2 and 3) (Table 1).

5.1 How to Meet the Challenges?

To meet the challenges in the real-time conditions our research adopts three techniques in camera technology. They are as follows (Figs. 4, 5 and 6):

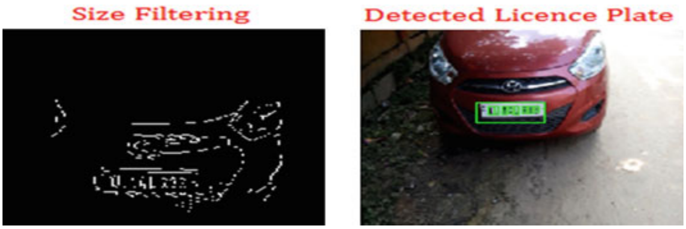


Fig. 2 Size filtering and genetic algorithm phase to detect the registration plate in MATLAB

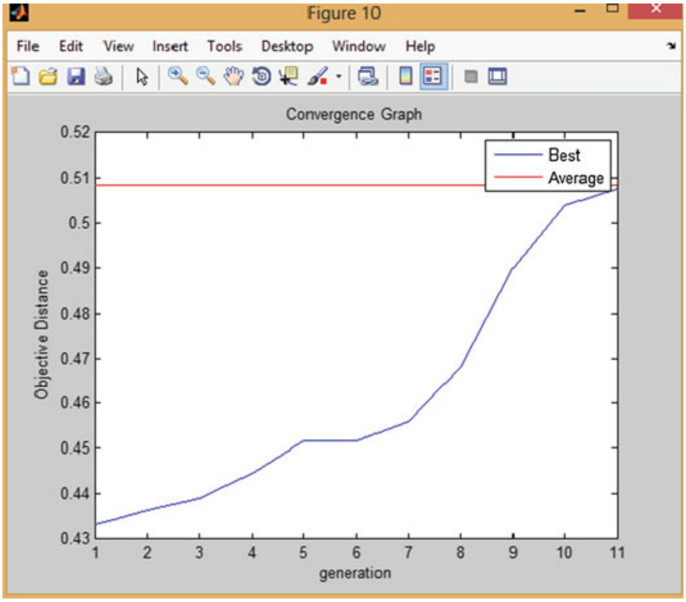


Fig. 3 Convergence graph plotted between objective distance and number of generations

Table 1 Confidence versus time comparison

Plate recognition	Confidence (%)	Time taken in (ms)
WOBVW14	98.77	50.43
WOBVV14	85.56	40.67
W0BVW14	81.65	30.53
W0BVV14	70.42	20.61

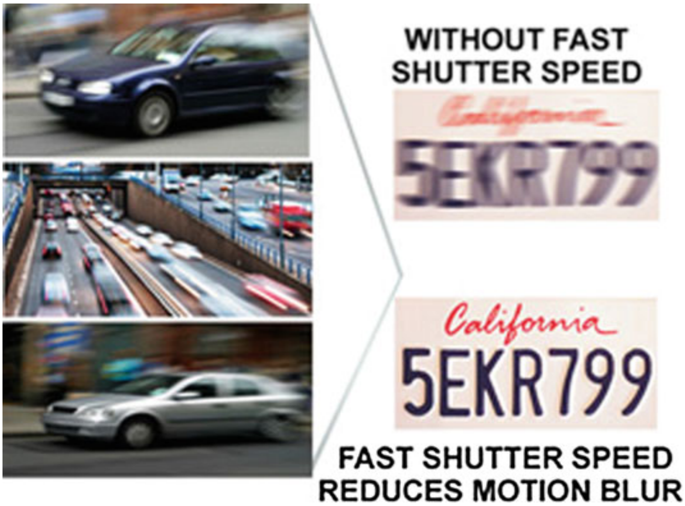

WITHOUT FAST SHUTTER SPEED
California
5EKR799
California
5EKR799
FAST SHUTTER SPEED REDUCES MOTION BLUR

Fig. 4 Example showing how fast shutter speed cameras reduce motion blur (http://www.licenseplatesrecognition.com/hardware-involved-in-lpr.html)

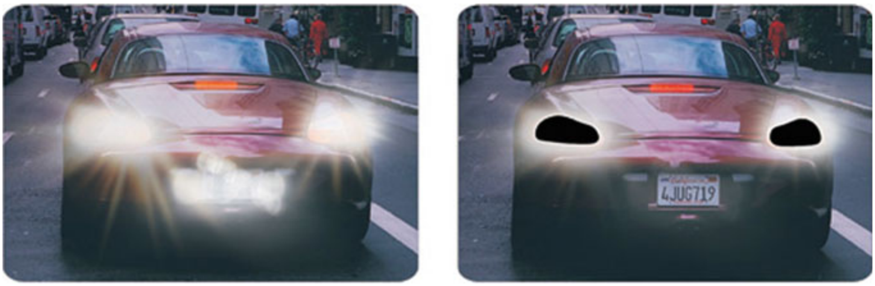


Fig. 5 Example of a light bending technology (http://www.licenseplatesrecognition.com/hardware-involved-in-lpr.html)

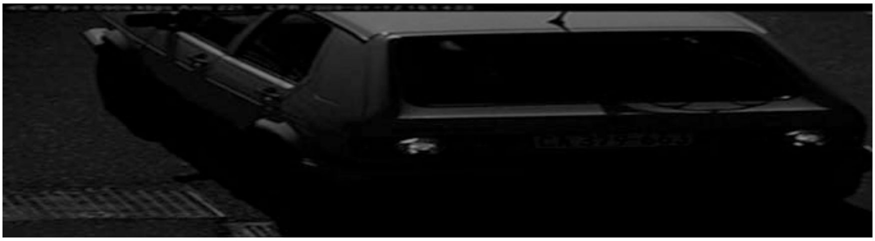


Fig. 6 Example showing infrared illumination technology (http://www.licenseplatesrecognition.com/hardware-involved-in-lpr.html)

- Fast shutter speed cameras
- Light bending technology
- Infrared illumination technology

6 Conclusion and Future Work

Our proposed method can be used for vehicle detection and tracking method based on multiple vehicle salient parts using a stationary camera. License plate can be considered as a salient feature for the tracking. In the future, localization of more salient parts other than the license plate of the vehicle such as the rear light, the front lights, the windshield, and front cover, aiming to deal with more severe occlusion and vehicle posture variation can be done for tracking the vehicle. In addition, vehicle detection and tracking process can be integrated into an embedded camera platform as a low-cost implementation.

References

1. "Vertical Edge Based Car License Plate Detection" Al-Ghaili, Syamsiah Mashood, Abdul Rahman Ramli and Alyani Ismail, IEEE Transactions On Vehicular Technology, Vol. 62, January 2013.
2. "Detecting License Plate Using Texture And Color Information", Ahmadyfard And Abolghasemi, In Proc. Int. Symp. Telecommun. 2008, Pp. 804–808. 978-1-4244-2751-2/08/2008 IEEE.
3. "Automatic License Plate Recognition", Shyang-Lih Chang, Li-Shien Chen, Yun-Chung Chung, And Sei-Wan Chen, Senior Member, IEEE IEEE Transactions On Intelligent Transportation Systems, VOL. 5, MARCH 2004.
4. "Car License Plate Detection Based On MSER", Wei Wang, Qiaojing Jiang, Xi Zhou And Wenyin Wan, School Of Telecommunications Engineering Xidian University 978-1-61284-459-6/11/2011 IEEE.
5. "An Online Self-Learning Algorithm for License Plate", Francisco Moraes Oliveira Neto, Lee D. Han, And Myong Kee Jeong, Senior Member, IEEE. IEEE Transactions On Intelligent Transportation Systems, Vol. 14, December 2013.

Comparing the Capacity, NCC, and Fidelity of Various Quantization Intervals on DWT

Velpuru Muni Sekhar, K.V.G. Rao, N. Sambasive Rao and Merugu Gopi Chand

Abstract A robust cover content extraction and embedding technique that trades off between visual quality and embedding capacity is proposed in this paper. In addition, adaptive quantization is used to achieve higher capacity of embedding with good visual quality. In this technique we are using Discrete Wavelet Transform (DWT) plus adaptive quantization to reduce noise over modification. Here, secret data is embedded into Non-zero quantized coefficients. By using this technique, we achieve approximately 0.99 Normalization Cross Coefficient (NCC) and Peak Signal-to-Noise Ratio (PSNR ≈ 60–70 dB). Comparison of results demonstrates that the proposed technique is better than the exiting techniques.

Keywords DWT · PSNR · Adaptive quantization · Data embedding and extraction and steganography · NCC

V.M. Sekhar (✉) · M.G. Chand
Vardhaman College of Engineering, Hyderabad, India
e-mail: munisek@gmail.com

M.G. Chand
e-mail: gopi_merugu@yahoo.comI

K.V.G. Rao
G. Narayanamma Institute of Technology & Science, Hyderabad, India
e-mail: kvgrao1234@gmail.com

N.S. Rao
Sumathi Reddy Institute of Technology for Women, Warangal, India
e-mail: snandam@gmail.com

H.S. Saini et al. (eds.), *Innovations in Computer Science and Engineering*,
Advances in Intelligent Systems and Computing 413,
DOI 10.1007/978-981-10-0419-3_6

1 Introduction

Nowadays every person getting services from digital computers. So, this era is called the digital era. The wide digitization of modern world helps fast economic growth and transparent governance of a country. Due to wide digitization, Internet becomes digital medium for data transmission [1, 2]. However, being a fully open medium, digitization brought us not only convenience but also security hazards. If data is transmitted through Internet, it gives convenience as well as risk. Some malicious users can create illegal copies or destroy or change the data in the Internet. It leads to security problems such spoofing, sniffing, man-in-the-middle, etc. [3]. These malicious attacks breach following security services such as confidentiality and authentication [4, 5]. It again creates inconvenience to the user. So, digitization not only brings advantages but also creates some challenges. To overcome these challenges, research community developed cryptographic and data hiding-based techniques [6]. Both have their own advantages and disadvantages. Cryptography provides security only between two end parties, once data is decrypted then no security to the content [4]. It cannot provide security to the broadcasted data. To provide authentication to the broadcasted content digital watermarking is used and to transfer secret data imperceptibly between two parties steganography is used [4].

In this paper we have focused on secret data transfer with data hiding technique, i.e., steganography.

1.1 *Steganography*

It is an art of science, which concealing a message into a cover content in imperceptible way. In Fig. 1 and Eq. 1 it demonstrates the process of data embedding [5, 7–10].

$$\mathrm{Stego}(i, j) = E(C(i, j), S(n)) \quad (1)$$

Fig. 1 Embedding secret media file

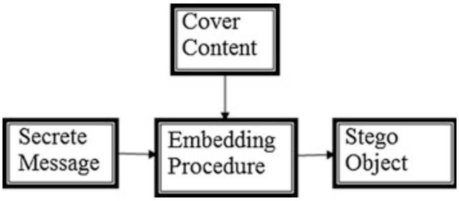
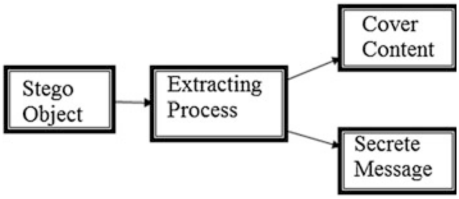


Fig. 2 Extraction of secret media file



Here, secret message '*S(n)*' is embedded in cover content '*C*(*i, j*)' where *n* denotes index of secret message bit, (*i, j*) denotes cover content position identifiers, *E* represents embedding process and Stego(*i, j*) represents output of embedding process.

Stego object encompasses both cover and secret message, but secret message cannot be intercepted by third party. The process of extracting secret message hidden in a stego object using steganography is called as *steganalysis* [10]. Figure 2 and Eq. (2) demonstrate the extracting process.

$$C^I(i,j) = Ex(\mathrm{Stego}(i,j), S(n)) \tag{2}$$

where *Ex* represents the extraction process and $C^I(i, j)$ represents output of extraction process.

Here, our objective is to minimize noise in stego image with maximum *n* value and extracted cover content should be similar to the original cover content ($C(i,j) \approx C^I(i,j)$). That can be achieved as follows, by applying DWT on cover content, followed by adaptive quantization on DWT coefficient and then embed the secret data *S* in quantized non-zero coefficients, following which data embedding can be performed.

1.2 Discrete Wavelet Transform

The wavelets transform grabs massive attention from research community, with effective transforming property [11–16]. It transforms cover content (digital image) from spatial domain to frequency domain. Various wavelets exist to perform transform, i.e., Haar, Symlet, Coiflet, and so on. In this paper, we are using 2D DWT Haar [17] wavelet transform. The 2-D Haar DWT is performed as follows: (i) Vertical Division Operation and (ii) Horizontal Division Operation [1].

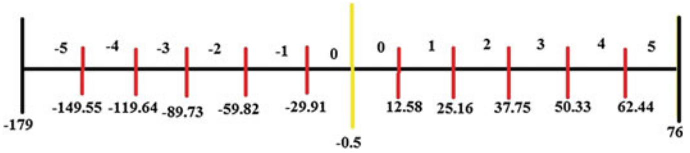


Fig. 3 Adaptive quantization structure

Horizontal division Operation: This division operation divides an image into vertically two parts. The first part is the sum of two adjacent columns, which is stored in the left side part as low-frequency coefficients. The other part is the differences of two adjacent columns, which is stored in the right side part as high-frequency coefficients. *Vertical Division Operation*: the division operation divides horizontally an image into two parts. The first part is the sum of two adjacent rows, which is stored in the upper side as low-frequency coefficients. The other part is the difference between adjacent rows, which is stored on the lower side as low-frequency coefficients, then all coefficient values are divided by 2.

Inverse DWT is same as inverse process of Vertical and Horizontal division operations.

1.3 Adaptive Quantization

It divides coefficients into variable quantized interval based on asymmetric characteristics of given coefficients [18–20]. Adaptive quantization encoding performs the following steps, i.e, [1].

(i) Find median ‘α’ from DWT coefficients ‘$H(i, j)$’ using Eq. (3) where ‘n’ denotes total number of coefficients

$$\alpha = \begin{cases} x_i \,|1 \le i \le n,\ i = \frac{n}{2} + 1\, x_1 \le x_2 \le \ldots \le x_n & \text{If}\, \frac{\text{n}}{2} \ne 0 \\ \frac{x_i + x_{i+1}}{2} \,|1 \le i \le n,\ i = \frac{n}{2}\ x_1 \le x_2 \le \ldots \le x_n & \text{If}\, \frac{\text{n}}{2} = 0 \end{cases} \tag{3}$$

(ii) Subtract median ‘α’ from coefficients ‘*H(i, j)*’

$$D(i,j) = H(i,j) - \alpha \tag{4}$$

(iii) Find right interval width ‘Δb_R’ and left interval width ‘Δb_L’ as shown in Fig. 3 using Eqs. (5) and (6). Let us assume left interval width = 29.41 and right interval width = 13.08.

$$\Delta b_L = \begin{cases} \frac{|\alpha - \min(H(i,j)|}{l/2}, \ \forall H(i,j) \in \{0, 1, 2 \ldots .255\} \end{cases} \tag{5}$$

$$\Delta b_R = \begin{cases} \frac{|\max(H(i,j) - \alpha|}{l/2}, \ \forall H(\,i,j) \in \{0, 1, 2 \ldots .255\} \end{cases} \tag{6}$$

(iv) Encode the coefficients '*H(i, j)*' using Eq. (7)

$$Q(i,j) = \frac{D(i,j)}{|D(i,j)|} \times \left\lfloor \frac{|H(i,j) - \alpha|}{\Delta b_L} \right\rfloor \tag{7}$$

$$Q(i,j) = \frac{D(i,j)}{|D(i,j)|} \times \left\lfloor \frac{|H(i,j) - \alpha|}{\Delta b_R} \right\rfloor \tag{8}$$

Reconstruction of adaptive quantization coefficients performs as follows [1]:

$$R_{Q(i,j)} = \begin{cases} (Q(i,j) + r) \times \Delta b_R + \alpha, & Q(i,j) > 0 \\ (Q(i,j) - r) \times \Delta b_L + \alpha, & Q(i,j) < 0 \\ 0, & \text{otherwise} \end{cases} \tag{9}$$

The rest of the paper is organized as follows. Section 2, discuss about literature review. Section 3, demonstrate embedding Technique. Section 4, discuss the comparison of results with some existing methods and proposed method with various intervals. Finally, Sect. 5, concludes the paper.

2 Literature Review

The data hiding in the cover content introduces some modifications to the original cover content [10–12]. The literature review schemes do not address trade-offs between embedding capacity and visual quality. There the trade-off between capacity and visual quality is not an important concern, rather their focus is only on individual requirements such as embedding capacity or visual quality or normalized cross correlation or robustness or fragility. But there are some applications such as medical imaging, military communication, fine arts, multimedia archive management, and remote sensing. These require the trade-off between capacity and visual quality [19]. The schemes which aim the above applications are referred as data hiding schemes [5, 7, 10, 15, 17, 19, 20].

Po-yueh chen et al. [18] have introduced the adaptive quantization paradigm for the first time. They showed that by applying quantization on frequency domain coefficients one can achieve better compression rate on digital images. Phadikar et al. [21] proposed a transform domain data hiding scheme for quality access control of images. The original image is decomposed into tiles by applying n-level lifting-based discrete wavelet transformation (DWT). A binary watermark image (external information) is spatially dispersed using the sequence of number generated by a secret key. The encoded watermark bits are then embedded into all DWT coefficients of nth level and only in the high-high (HH) coefficients of the subsequent levels using either modulation (DM) but without complete self-noise suppression.

Most of the existing schemes on data hiding exist for spatial domain Least Significant Bit (LSB) embedding technique and Multiple LSB embedding techniques [5] or frequency domain data embedding [7] or frequency domain and uniform quantization technique [21]. To the best of our knowledge, no data hiding scheme exist on DWT plus adaptive quantized nonzero coefficients data embedding.

3 Proposed Data Hiding Technique

In this section, we deals with embedding algorithm and extraction algorithm of the proposed Data hiding/steganography scheme. In Sect. 1 illustrates the basic information about steganography technique that comprises of three modules: DWT, adaptive quantization, and embedding and extraction. The steganographic embedding process is designed to be performed in the DWT domain. This has several advantages. DWT is used in the most popular image and video compression formats, including JPEG, MPEG, etc. Adaptive quantization is performed on DWT coefficients to further compress the images and videos. After DWT and adaptive quantization obtained the coefficient consists of zero and Non-zero values [19]. These Non-zero values are embedded with secret message as shown in Eq. (10).

$$e = \begin{cases} \frac{c}{|c|} \times \text{floor}((2\log_2(2|c|))) & \textit{if } S(i) = 0 \\ \frac{c}{|c|}\text{floor} \times (2\log_2(2|c|) - 1) & \textit{if } S(i) = 1 \end{cases} \tag{10}$$

where '*c*' is quantized Non-zero coefficients, '*i*' is *i*th data bit in secret message and '*e*' is the modified version of *c*.

Extraction of secrete message as shown in Eq. (11).

$$S(i) = \begin{cases} 0 & if\ \frac{e}{2} = 0 \\ 1 & \text{otherwise} \end{cases} \qquad (11)$$

Proposed Embedding and Extracting algorithms

Embedding Algorithm

```
1. Read the cover image C= {c(I,1),c(1,2),.,c(512,512) }.
2. Calculate Diagonal (HH), Vertical (LH), Horizontal (HL) and
   Approximation (LL) coefficient to the cover content 'C' by applying
   DWT.
3. Find the displacement matrix D for all sub-bands.
      i) Select a sub-band and identify the median 'ω' using equation 3
      ii) Subtract the 'ω' median with all sub-band coefficients as
   like equation 4.
4. Select the left and region intervals for all sub-bands using
   equation 5 and 6.
5. Calculate the adaptive quantization matrix for HH, HL and LH sub-
   bands using equation 7 and 8
6. Calculate the embedding matrix for HH, HL and LH sub-bands using
   equation 10.
7. Reconstruct the image '$C^I$' using reverse process of above and
   inverse DWT.
```

Extracting Algorithm

```
1. Read the stego image $C^I$= {$C^I$(1,1),$C^I$(1,2),…., $C^I$(512,512) }.
2. Calculate Diagonal (HH), Vertical (LH), Horizontal (HL) and
   Approximation (LL) coefficient to the stego image '$C^I$' by applying
   DWT.
3. Calculate the extracting matrix and retrieving secrete data from HH,
   HL and LH sub-bands using equation 11.
4. Reconstruct the image using reverse process of above and inverse
   DWT.
```

4 Result Analysis and Comparisons

A set of JPEG images, of size 512x512 pixel was used for the experiment is Aerial, Airplane, Baboon, Barbara, Boat and couple. The secret data is generated by using '*rand'* function in MATLab by the '*rand(1*, *count)>0.5*.' It generates a '*count-by-1*' column vector containing 0 or 1 number drawn from a uniform distribution. Here "*count*" defines number of Non-zero coefficients after DWT plus Adaptive

Quantization. The Data hiding scheme in various is compared with the following parameters: Capacity, quantization intervals, PSNR values.

Table 1 Embedding capacity (EC), PSNR and NCC in different quantization intervals

Intervals	Aerial			Airplane			Baboon			Barbara		
	EC	PSNR	NCC	EC	PSNR	NCC	EC	PSNR	NCC	EC	PSNR	NCC
4-4-4	93663	32.43	0.9797	73649	39.52	0.9855	98397	32.32	0.9395	95345	36.66	0.9522
8-8-8	98696	42.85	0.9872	74423	51.71	0.9954	106986	42.18	0.9645	103727	46.62	0.9724
16-16-16	109084	52.59	0.9921	75700	64.07	0.9985	124551	50.89	0.9761	114543	56.66	0.9842
32-32-32	124130	61.94	0.9964	78319	75.11	0.9999	145268	59.87	0.9913	128267	66.84	0.9971
64-64-64	140652	66.44	0.9971	86379	78.69	1.0003	164661	62.74	1.009	149047	66.38	1.0051
128-128-128	155896	62.62	0.9849	104120	76.84	0.9998	179079	57.38	1.0118	170144	60.58	1.0083

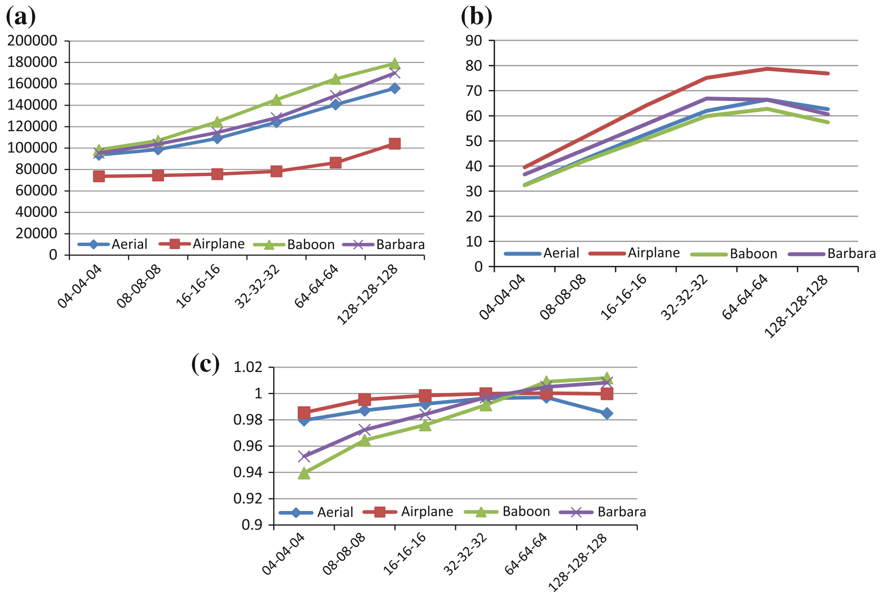


Fig. 4 Quantization intervals versus EC, PSNR, and NCC

Table 2 Po-Yueh chen [5] data embedding
(a) Embedding Capacity and
(b) Visual quality

Images	Cases		
	Case 1	Case 2	Case 3
	Capacity	Capacity	Capacity
Airplane	376710	507856	573206
Baboon	376835	507670	573392
Boat	376598	507867	573318
Girl	377038	507940	573422
Lena	376942	507856	573550
Pepper	377125	507946	573184
Images	Cases		
	Case 1	Case 2	Case 3
	PSNR	PSNR	PSNR
Airplane	50.8554	45.9961	44.7683
Baboon	50.7647	46.1948	44.9664
Boat	50.7499	46.1385	44.9260
Girl	50.7746	46.0763	44.8842
Lena	50.8021	46.0882	44.9011
Pepper	50.7975	46.0793	44.8973

Table 3 Sagar [19] data embedding scheme

Image name	DCT		
	Capacity	PSNR	NCC
Aerial	56564	24.2961	0.9756
Airplane	12608	31.6485	0.9942
Baboon	74017	22.3393	0.9664
Barbara	39346	25.8765	0.9824

Table 1 and Fig. 4 present a clear idea of how embedding capacity (EC), PSNR, and NCC vary in quantization intervals. In which interval does the maximum embedding capacity exist? What is the range in an interval for different input images of embedding capacity? In 128-128-128 interval for input image embedding capacity lies between [≈100,000, ≈170,000] for different input images. Average embedding capacity value around 140,000 in interval 128-128-128 for different input images. What is the approximation value in an interval for different input images of NCC? In interval 128-128-128 for different input images is NCC approximately ≃1. What is the range in an interval for different input images of PSNR? In 128-128-128 interval for input images range is [≈25, ≈80]. Average PSNR value around 65 for 128-128-128 interval for different input images. Comparison of the proposed method results with exist methods Tables 2 and 3.

5 Conclusion

Proposed data embedding scheme is compared with existing schemes and also within the scheme for different intervals with respect to embedding capacity and PSNR. Based on various user performance measures, several intervals are provided for selection. According to the analysis and comparison results, the PSNR is still a satisfactory value and even the highest embedding capacity is noticed. This is due to different characteristics of DWT coefficient in different sub-bands. Here the sensitive and essential (the low frequency part) part is kept unchanged while embedding secret data. From this it is inferreds that this scheme performs better than existing schemes.

References

1. European Commission report on the application of the IPR Enforcement Directive (2004/48/ec) http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2010:0779:FIN:EN:PDF (last visited 4 March 2015).
2. Muni Sekhar V, K. Venu gopal Rao and N. Sambasiva Rao (2011) Convention Cloud application development: SOA. In: International Journal of Advanced Computing Vol. 3(3) pp 108–112.
3. Muni Sekhar V, K. Venu gopal Ra, N. Sambasiva Rao and M. Ravi Kumar (2013) Guaranteed Quality of Service in cloud ready application. In: IEEE conference of International Symposium on Computational and Business Intelligence 2013, New Delhi, India pp 24–28.

4. Ingemar J. Cox, Mathew L. Miller, Jeffrey A. Bloom, Jessica Fridrich and Ton Kalker (2008) Digital Watermarking and Steganography. As a text book: 2nd Edition MK publication, 2008.
5. Po-Yuch Chen and Hung-Ju Lin (2006) A DWT Based Approach for image steganography.In: International Journal of Applied Science and Engineering 2006, Vol. 4(3) pp 275–290.
6. Asoke K Talukder, Muni Sekhar V, Vineet Kumar Maurya, Santhosh Babu G, Jangam Ebenezer, Jevitha K P, Saurabh Samanta, Alwyn Roshan Pais (2009) Securityaware Software Development Life Cycle (SaSDLC)–Processes and Tools. In: IEEE Wireless Optical Computer Networks 2009. WOCN'09, Cairo pp 28–30.
7. Ayman Ibaida and Ibrahim Khalil (2013) Wavelet-Based ECG Steganography for Protecting Patient Confidential Information in Point-of-Care Systems. In: IEEE Transactions on Biomedical Engineering. Vol. 60 (12), pp 3322–3330.
8. Neil F. Johnson, Sushil Jajodia (1998) Exploring Steganography: Seeing the Unseen. In: IEEEComputer Practices report, George Mason University, Vol. 31 (2), pp 26–34.
9. Ross J. Anderson and Fabien A. P. Petitcolas (1998) On the Limits of Steganography, IEEE Journal on Selected Areas in Communications, Vol. 16 (4) pp 474–481.
10. Muni Sekhar V, Naresh Goud M and Arjun N (2014) Improved Qualitative Color Image Steganography based on DWT. In: International Journal of Computer Science and Information Technologies, Vol. 5 (4), pp 5136–5140.
11. B.N. Chatterji, Manesh Kokare, A. Adhipathi Reddy, Rajib Kumar Jha (2003) Wavelets for Content Based Image Retrieval and Digital Watermarking for Multimedia Applications. In: IEEE Conference on Electronics, Circuits and Systems, Sharja, UAE, December 2003, pp. 812–816.
12. F. Kumgollu, A Bouridane, M A Roula and S Boussaktd (2003) Comparison of Different Wavelet Transforms for Fusion Based Watermarking Applications. In: IEEE conference on Electronics, Circuits and Systems, Sharja, UAE, December 2003, pp. 1188–1191.
13. Barni M, Bartolini F, Piva (2001) an Improved Wavelet Based Watermarking through Pixelwise Masking. In: IEEE transactions on image processing, Vol. 10(5), pp. 783–791.
14. Victor V., Guzman, Meana (2004) Analysis of a Wavelet-based Watermarking Algorithm. In: IEEE Proceedings of the International Conference on Electronics, Communications and Computer 2004, Mexico, pp. 283–287.
15. N. Kaewkamnerd and K.R. Rao (2000) Wavelet Based Image Adaptive Watermarking Scheme. In: IEEE Electronic Letters, Vol. 36(4), pp. 312–313.
16. Ester Yen, Kai-Shiang Tsai (2008) HDWT-based Grayscale Watermark for Copyright Protection. In: Expert Systems with Applications 35(2), Elsevier, August 2008, pp. 301–306.
17. Muni Sekhar V, K. Venu gopal Rao, N. Sambasiva Rao (2015) Enhanced Adaptive Data hiding in DWT. In IOSR Journal of Computer Engineering Vol. 17(2) pp 30–40 DOI: 10.9790/0661-17263040.
18. Chen, P. Y., Wu, S. H., and Chen, G. F. (2007) An Adaptive Quantization Scheme for JPEG2000. In: 6th Symposium on Management of Information Technology and Personal Training, Taipei.
19. Sagar Gujjunoori and B. B. Amberker (2013) A DCT Based Near Reversible Data Embedding Scheme for MPEG-4 Video. Proceedings of the Fourth International Conference on Signal and Image Processing, Lecture Notes in Electrical Engineering Volume 221, 2013, pp 69–79.
20. Phadikar, A., Maity, S. P., and Kundu, M. K. (2008) Quantization Based Data Hiding Scheme for Efficient Quality Access Control of Images Using DWT via Lifting. In: IEEE Sixth Indian Conference on Computer Vision, Graphics & Image Processing 2008, pp 265–272.
21. Phadikar, A., Maity, S. P., and Kundu, M. K. (2008) Quantization Based Data Hiding Scheme for Efficient Quality Access Control of Images Using DWT via Lifting. In: 6th ICVGIP'08, Bhubaneswar, pp. 265–272.

Enhancing Cloud Computing Server to Use Cloud Safely and to Produce Infrastructure of High Performance Computing

Kavita Sultanpure and L.S.S. Reddy

Abstract Significance of cloud computing is getting popular in industrial and scientific communities for its numerous advantages and it is not free from its drawbacks. However, it has been observed that major issues such as security, compliance, legal, and privacy matters relative to risk areas like external storage of data, shortage of control, public Internet dependency, integration and multitenancy with internal security are not addressed to their fullest extent. Conventional security mechanisms like authorization, authentication, and identity are found to be inadequate for present cloud users. Moreover, controls of security in cloud computing are unique than controls of security in any information technology environment with respect to deployment, technologies, and operations. Therefore, in order to address the aforesaid issues this research intends to focus on enhancing the cloud computing server for security, performance, and load balancing to use cloud safely for high performance computing and to produce infrastructure of high performance computing.

Keywords Cloud computing · High performance computing · Security

1 Introduction

Cloud computing is a revolutionary technology aiming at giving different storage and computing services over the Internet. Cloud computing includes software, platform, and infrastructure as services. Service providers of cloud rent data center software and hardware for delivering computing and storage services via Internet. Internet users could able to retrieve cloud services by adopting cloud computing as

K. Sultanpure (✉) · L.S.S. Reddy
CSE Department, K L Education Foundation, Vaddeswaram, A. P., India
e-mail: kavita.sultanpure@gmail.com

L.S.S. Reddy
e-mail: drlssreddy@kluniversity.in

H.S. Saini et al. (eds.), *Innovations in Computer Science and Engineering*,
Advances in Intelligent Systems and Computing 413,
DOI 10.1007/978-981-10-0419-3_7

if they were deploying on super computer. They could store their information in the cloud rather than on their own devices, to access ubiquitous data more possible [1]. According to Buyya et al. [2] cloud computing gives an access for computer user for services of information technology such as servers, applications, storage of data without a necessity of the technology or infrastructure ownership. Cloud computing is a model to enable on-demand network and convenient access to a shared pool of configurable computing resources.

Cloud computing seems to be distribution architecture and computation paradigm and it is intended to give quick, secure, and convenient storage of data and service of net computing, with all resources of computing viewed as services and delivered via Internet [3, 4]. Moreover, cloud improves agility, ability, availability, collaboration, and scalability for adapting to fluctuations based on demand, accelerate work of development, and gives capacity for reducing the cost through efficient and optimized computing [5–7]. High performance computing (HPC) permits engineers and scientists to solve complex engineering, science, and business issues using applications that need enhanced networking, high bandwidth, and very high capabilities for computation.

Security of data on the side of cloud is not concentrated on the transmission of data process, but also security in the system and protection of data for that information stored on cloud side storages. From the security of data perspective, which was considered as significant aspect of service quality, cloud computing inevitably poses many threatening security issues for numerous reasons as storage of data in the cloud would be updated by clients encompassing deletion, reordering, appending, modification, insertion, and so on. This fact necessitates mechanism to facilitate correctness of storage under dynamic data upgrade. Thus, distributed protocols for assurance of storage correctness would be more significant to achieve a secure and robust storage system in the cloud data in real world [8].

In this view, we propose the *in memory cache* to improve the security in the server of cloud computing to use cloud safely for high performance computing as well as to produce infrastructure of high performance computing.

2 Related Work

According to Aljaber et al. [9] storage of multimedia file in cloud computing needed the security. Multimedia cloud computing is referred as multimedia computing over grids, network of content delivery is adopted for minimizing the latency and maximize the data bandwidth, server-based computing, and so on. It provides infrastructure for high performance computing aspect. Youssef and Alageel [10] have proposed a generic cloud computing security model which assists to fulfill privacy and security needs in the clouds and safeguard them against different vulnerabilities. Talib et al. [11] has examined the issue of data security in the environment of cloud computing, to ensure the correctness environment, confidentiality, integrity, and availability of client's data in the cloud. To achieve this security framework a multiagent system to provide security of cloud data storage was developed. Further Rasch model was adopted to analyze pilot questionnaire. Reliability of item is identified to be poor, a few items and participants were found as misfits with distorted estimations.

Suganya and Damodharan [12] have carried out an investigation to enhance security in cloud computing for storage services. This research discussed about issue in the security of data and storage of data in cloud is basically a distributed system of storage. For achieving guarantee of data availability and integrity in the cloud and for enforcing quality of service reliable storage in the cloud to users or clients, distributed one was proposed in a flexible and efficient scheme. Kaur and Singh [13] have implemented encryption algorithms for enhancing security of data of cloud in cloud computing. This study discussed about the Cipher cloud. Such framework let client to keep the information very confidentially in the public cloud. Apart from these, it was noted that security controls required to safeguard most sensitive information would not be guaranteed in architectures of public cloud computing.

The research by Guleria and Vatta [14] focused on enhancing multimedia security in cloud computing environment using algorithm of crossbreed. This study develops a more flexible and effective distributed scheme of verification for addressing the security issue in the storage of data in cloud computing. Apart from these, it was noted that there is feeling that security associated with regulatory compliance is issue in adopting cloud computing. Sachdev and Bhansali [15] have enhanced the security of cloud computing using Advanced Encryption Standard (AES) algorithm. This study developed a simple model for protection of data where data is encrypted using AES prior it is deployed in the cloud, therefore ensures data security and confidentiality. Ukil et al. [16] have proposed architecture and analysis to incorporate unique security protocols, techniques, and schemes for cloud computing especially in IaaS (Infrastructure-as-a-service) and PaaS (Platform-as-a-service) systems. Sandhu and Singla [17] has developed an approach to enhance the security of model technology of multimedia data based on cloud computing. This research develops a more flexible and effective distributed scheme of verification for addressing the storage of data security problem in cloud computing.

3 Proposed in Memory Cache System

As highlighted in reviewed literature, the conventional setup of cloud servers ignores the aspects of load balancing, security, and speed pertaining to requesting applications. In order to tackle these crucial issues, we have proposed the inclusion of *In Memory Cache* in cloud servers for high performance computing as depicted in Fig. 1.

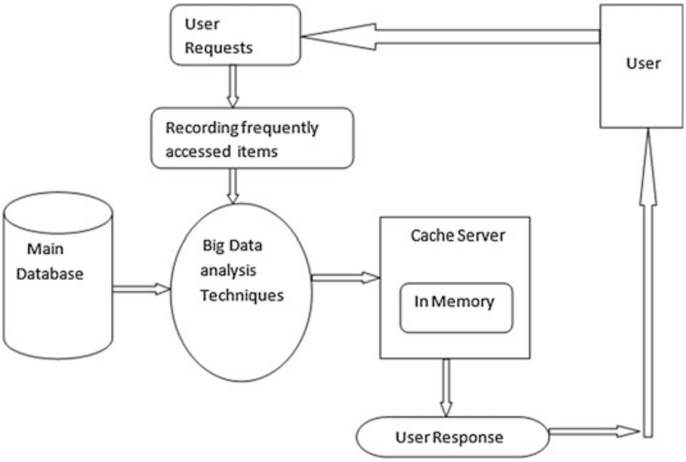


Fig. 1 *In memory cache* server in Cloud Server for HPC

In conventional cloud-based setup, there is involvement of all the servers while responding to the user queries. However, doing the same rigorous procedure for the same pattern of queries is not only time consuming but also increases processing overheads of the servers. In this view, inclusion of *In Memory Cache* is suggested to store the frequently searched items. It responds to the requests without disturbing main servers. Moreover, the developed cache is secure as the dependency on the whole server database is removed and all the responses will be generated from temporary database. Consequently, enhanced speed and less load on the server are observed.

We have explained proposed architecture using case study of access pattern of electronic commerce. Let us assume that an online electronic commerce store is running in a server of cloud. If, user visits that site and transmits requests and server responds to the request, at that time, if electric commerce admin wants that information, then it is possible to retrieve some data from cache memory. *In memory cache* will store some temporary information as per frequent requests in the cache server. It might be any serial key or authentication key or any general data. In such case, web services API particularly SOAP will assist us to link it from the database of cache. Big Data analysis technique using Hadoop can be used to analyze those frequently searched items. This *in memory cache* server can be integrated with Amazon.

Main purpose of the study is to develop that structure of data in the server and implement the web service API. Finally, by adopting our product anybody can embed it with their products or services wherever it is necessary.

4 Conclusion

The proposed *In Memory Cache* for cloud architecture overcomes the limitations of conventional setup by providing high performance, speed, and security in computing. *In Memory Cache* stores the frequently searched items and responds to the requests without disturbing main servers. Consequently, enhanced speed and less load on the server are observed. With this we can able to retrieve some essential data from the system of cache instead of retrieving the whole storage. It will maximize the security of the server. This setup may be integrated with the Amazon.

References

1. Gunho Lee et al, "Above the Clouds: A Berkeley View of Cloud Computing," Engineering and Computer Sciences University of California at Berkeley Technical Report No. UCB/EECS-2009-28, February 10, 2009.
2. Rajkumar Buyya, Chee Shin Yeo, Srikumar Venugopal, "Market-oriented CC: Vision, Hype and Reality for Delivering it services as Computing utilities," in:Proc.10th IEEE International Conference on High Performance Computing and Communications, HPCC 2008, Dalian, China, Sept. 2008.
3. Zhang S, Chen X, Huo X, "Cloud Computing Research and Development Trend," In: Second International Conference on Future Networks (ICFN'10), Sanya, Hainan, China. IEEE Computer Society, Washington, DC, USA, 2010, pp 93–97.
4. Zhao G, Liu J, Tang Y, Sun W, Zhang F, Ye X, Tang N, "Cloud Computing: A Statistics Aspect of Users," In: First International Conference on Cloud Computing (CloudCom), 2009, Beijing, China. Springer Berlin, Heidelberg, pp 347–358.
5. Marinos A, Briscoe G, "Community Cloud Computing," In: First International Conference on Cloud Computing (CloudCom), 2009, Beijing, China, Springer Berlin, Heidelberg.
6. Cloud Security Alliance (2011) Security guidance for critical areas of focus in Cloud Computing V3.0. Retrieved on: 23rd May 2015, https://cloudsecurityalliance.org/guidance/csaguide.v3.0.pdf.
7. Khalid A, "Cloud Computing: applying issues in Small Business," In: International Conference on Signal Acquisition and Processing (ICSAP'10), 2010, pp 278–281.
8. Cong Wang, Qian Wang, Kui Ren,Wenjing Lou, "Ensuring Data Storage Security in Cloud Computing," 17th international workshop on Quality of Service, 2009, IEEE, vol 186, no 978, pp 1–9.
9. B. Aljaber, T. Jacobs, K. Nadiminti, R. Buyya, "Multimedia on global grids: A case study in distributed ray tracing," Malaysia. Journal Computer Science, vol. 20, no. 1, pp 1–11, June 2007.
10. Youssef. E, Alageel. M, "A Framework for Secure Cloud Computing," International Journal of Computer Science, vol 9, iss 4, 2012.
11. Talib. A et al, " Towards a Comprehensive Security Framework of Cloud Data Storage based on multi-agent system architecture," Journal of information security, vol 3, No 4, 2012, pp 295–306.
12. Suganya. S and Damodharan. P, "Enhancing Security for Storage Services in Cloud Computing," International Journal of Scientific and Research Publications, vol 3, iss 6, 2013, pp 1–3.

13. Kaur. M and Singh. R, "Implementing Encryption Algorithms to Enhance Data Security of Cloud in Cloud Computing," International Journal of Computer Applications, vol 70, no 18, May 2013.
14. Guleria. S and Vatta. S, "To Enhance Multimedia Security in Cloud Computing Environment Using Crossbreed Algorithm," International journal of application or innovation in engineering and management, vol 2, iss 6, 2013, pp 562–568.
15. Sachdev. A and Bhansali. M, "Enhancing Cloud Computing Security using AES algorithm," International Journal of Applications, vol 67, no 9, 2013.
16. Ukil. A, Debasish Jana, Ajanta De Sarkar et al, "A Security Framework in Cloud Computing Infrastructure," International Journal of Network Security and its applications, vol 5, no 5, Sep 2013, pp 11–24.
17. Sandhu. S and Singla. S, "An Approach to Enhanced Security of Multimedia Data Model Technology based on cloud computing," International Journal of Advanced Research in Computer Science and Software Engineering, vol 3, issue 7, 2013.

3-D Array Pattern Distortion Analysis

N. Dinesh Kumar

Abstract Indian MST Radar antenna comprises of 1024 three-component Yagi-Uda receiving antennas masterminded in a 32 × 32 matrix over a territory of 130 m × 130 m. The antenna array is fed by 32 distributed transmitters whose peak output power varies from 15 to 120 kW. Because of different reasons a few number of transmitters are nonoperational making the linear subarray relating to these transmitters ineffectual. Also, inside a subarray, it is conceivable that a few components would not get the excitation motion because of the detached association or discontinuity issues in the feeder line. The paper talks about these outcomes in the thinning of the aperture, array pattern distortion, and the deviation of the excitation from the predefined Taylor distribution. The constantly expanding demand in the advancement of software for aperture thinned radiation pattern has spurred to model the present work. Matlab is used to perform the investigation and to plot the radiation designs in both principle planes and in three-dimensional view.

Keywords 3-D array pattern · Aperture thinning · Array distortion · E-plane pattern · Matlab · Radiation pattern · Side lobes · Taylor distribution

1 Introduction

The work exhibited in the paper is developed at National MST Radar Facility (NMRF) (13.5°N, 79.2°E), an independent association under Department of Space (DOS) in the field of Atmospheric Science and Research, located at Gadanki on Chittoor highway, Chittoor District, Andhra Pradesh, India. This office has an obligation of leading examinations with radar for different investigative recommendations got from all over the nation in the field of atmospheric science and is

N. Dinesh Kumar (✉)
Vignan Institute of Technology & Science, Deshmukhi, India
e-mail: dinuhai@yahoo.co.in

H.S. Saini et al. (eds.), *Innovations in Computer Science and Engineering*,
Advances in Intelligent Systems and Computing 413,
DOI 10.1007/978-981-10-0419-3_8

likewise having an offline facility for examining such information got by directing the experiments [1].

Indian MST Radar antenna comprises of 1024 three-component Yagi-Uda receiving antennas [2] masterminded in a 32 × 32 matrix over a territory of 130 m × 130 m. 0.7λ (λ being the radar wavelength) spacing is maintained between interelements and is utilized as a part of both principal directions, which permits a grating lobe-free beam scanning up to a point of around 24° from broadside direction. 32 transmitters of varying power, each feeding a linear subarray of 32 elements, illuminate the array.

Due to various reasons a few number of transmitters are nonoperational making the linear subarrays corresponding to these transmitters ineffective. During the maintenance period, discontinuities are observed in the feeder lines. Even if the transmitters are operational, within a subarray, it is possible that some elements will not get the excitation signal [2]. This results in the thinning of the aperture and the deviation of the excitation from the specified Taylor distribution. Due to this deviation, the array pattern will get distorted.

2 Feeder Network Configuration

The feeder arrangement of MST radar antenna group contains two orthogonal sets; one for each polarization. The feeder framework embodies 32 parallel runs of *center-fed-series-feed* (*CFSF*) structures. Thirty-two transmitters of contrasting power edify the show; each supporting direct subarrays of 32 antenna components. The feeder frameworks of all the subarrays are indistinguishable to the degree that the power dissemination is concerned. The CFSF framework, including power divider at the point of convergence and a plan of directional couplers on each one side of it, associate the linear subarray to the TIR switch, which passes on the transmitter output power to the group and the power got by the array to the relating low-noise amplifier. Transmitter yield power passed on by the rigid connection (RG 1-5/8″) is brought to the data of the subcluster where the 3 dB power divider divides it into two proportionate measures for center feeding purpose. The partitioned powers are passed through two indistinguishable sets of directional couplers (each one involving 15 couplers laid in the arrangement). As the power goes through, every coupler passes on power to one radio antenna through its coupled port. The coupling components of the couplers are formed to get the modified Taylor distribution. The aperture excitation conveyance over the array takes after an estimate to altered Taylor weighing in both the principle directions. The coveted power appropriation across the array or cluster is proficient in one principle direction (E-plane) by the differential powers of the transmitters and in the H-plane by the suitable coupling coefficients of the CFSF system [1]. The measuring capacity was landed to realize a side lobe level (SLL) of −20 dB.

3 Implementation Methodology

3.1 Side Lobe Level Requirements

For getting the craved side lobe characteristics, the adequacy over the whole opening must be suitably decreased. It is decently understood that if one tries to diminish the SLL, the gap viability furthermore the gain diminish broadly. Consequently, to get the point by point increase one needs to trade off between the decrease capability and the peak SLL particularly. As indicated by the points of interest, at the MST radar center they picked −20 dB for SLL with decrease effectiveness of 80 % [1]. For the goal to be satisfied, modified Taylor distribution is chosen.

3.2 Modified Taylor Distribution

This distribution is utilized for MST radar applications as a decent comprise for achieving the main lobe and the side lobe specifications as set down. The mathematical function describing this amplitude distribution [3] is

$$A(Z) = \left(\frac{1}{2\pi}\right) J_0\left(i\pi B\sqrt{\left(1 - \frac{4Z^2}{d^2}\right)}\right) \tag{1}$$

where

J_0 is the Bessel function of the first kind

Z varies from $-d/2$ to $+d/2$

d is the total length of the aperture

B determines the level of the first SLL

To decrease the unpredictability of the feed network, certain measure of stepping is carried out on the constant gap appropriation. This quantization of the genuine decrease capacity prompts a viable aperture distribution. With this new measured aperture distribution the radiation sample of E-plane is plotted and computed using Matlab. This pattern reveals the following facts [4, 5]:

1. There is a 0.4 dB rise in the first side lobe level (SLL).
2. Far off, SLL does not decrease as quickly as they accomplish for a continuous distribution.
3. Around 45° the SLL rises roughly by 10 dB over the level computed for the continuous distribution.
4. No change in 3 dB beam width.

3.3 Computation of Far-Field Pattern

In the far-field, the electromagnetic waves discharging from the antenna components spread parallel to one another. The field strength [6] at any point due to an array of $2n$ elements is given by

$$E(\theta) = \sum_{n=1}^{2N} \frac{A_n \exp(-jkr_n)}{r_n} \quad (2)$$

where A_n = excitation current coefficient for the nth element, $k = 2\pi/\lambda$, phase constant, r_n = distance of the field point from the nth element, n = element number.

4 Results

Examination of Yagi-Uda antenna [2] has been overseen in point of interest close by the aperture thinning of MST Radar, and a straightforward easy to understand computer program is delivered to perform the examination and further to plot the radiation design in both standard planes. Aperture thinning package analysis is developed in Matlab with different menu options for easy analysis. The radiation design in both the fundamental planes can be plotted [4, 5]. Examination of Yagi-Uda antenna is executed and is extended to the advancement of thinned radiation pattern [7] of the Indian MST radar reception demonstrated. Matlab was picked for the numeric retribution and visualization of the output.

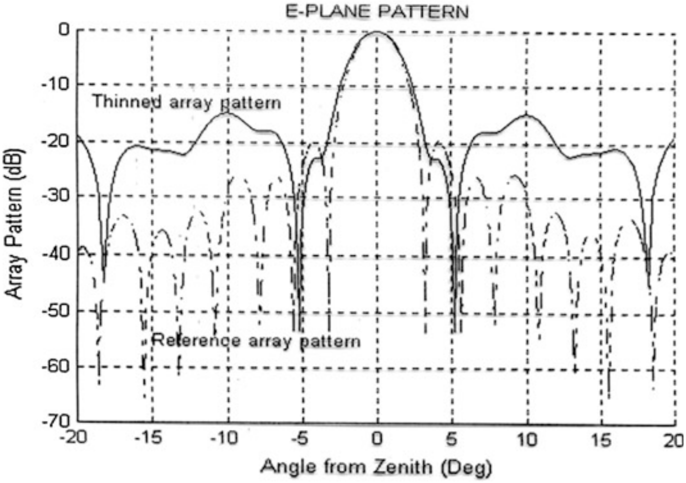


Fig. 1 E-plane pattern distortion due to array thinning radar antenna array

Figure 1 shows the radiation design for the fundamental and first few side lobes and is fundamentally same as the array design, subsequently the three-segment Yagi antenna pattern is truly expansive with 3 dB beam widths (BW) of around 66° in the E-plane and 120° in the H-plane. The array pattern shows that for the aperture distributed embraced an addition of 37 dB gain, a 3 dB BW of 2.65°, and a SLL of 20 dB could be made sense of. The 3-D power distribution for 32 antenna element array at MST radar facility to visualize the radiation pattern for the current power levels of the entire transmitter is switched on as shown in Fig. 2.

Normalized voltage levels are considered in the plots drawn. If transmitters 1–8 and 25–32 are OFF as shown in Fig. 3, we can clearly notice that 3-D power distribution has only few side lobes. A complete 3-D array pattern for the MST radar antenna array is shown in Fig. 4.

In Matlab package aperture thinning analysis menu “Status” option is available which allows the user to see the contents of the selected file having the array of transmitters. It also contains a submenu “Entry” that allows the user to enter the option for both transmitter power and other parameter values for further processing. The transmitter power if entered as 0, will give an indication that the particular transmitter is in OFF state. If the power level is beyond 120, it gives an error message and its value will be set to zero and edit box will get to red color. Once the values are entered and if “ACCEPT” button is pressed then the entered values will be processed.

Fig. 2 3-D power distribution for MST transmitters OFF: 6, 8, 9, 17, 29, 30, and 32, tilt of 0°

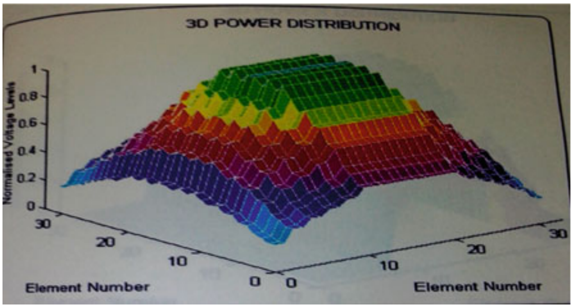
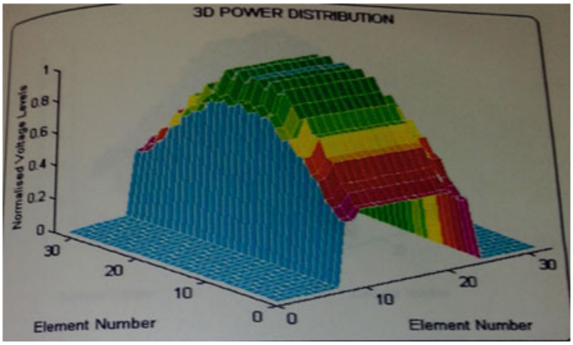


Fig. 3 3-D power distribution if transmitters 1–8 and 25–32 are OFF



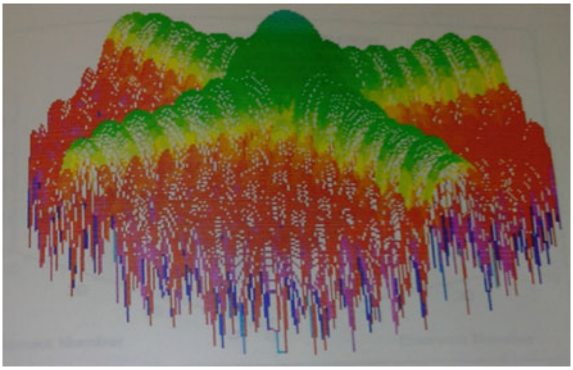


Fig. 4 MST radar antenna 3-D array pattern

Table 1 Variation of parameters with different array thinning configurations

S. no.	Subarray OFF	Loss in gain (dB)	SLL (dB)
1	Nil	0.00	−20.0
2	3, 4	0.40	−20.0
3	5, 6	0.50	−22.0
4	3, 4, 5, 6	1.00	−19.5
5	5, 6, 27, 28	0.55	−17.5
6	1–8	2.00	−16.0
7	1–8 and 25–32	4.00	−13.5
8	9,10	0.50	−18.3
9	11, 12	0.80	−17.7
10	13,14	0.80	−15.2
11	15,16	0.80	−14.2
12	13, 14, 15, 16	1.90	−11.5
13	15, 16, 17, 18, 19	2.00	−09.0
14	9–16	3.30	−07.0

5 Conclusion

A point-by-point examination of Yagi-Uda antenna apparatus nearby the aperture thinning of MST radar and a simple to utilize computer programs are made to perform the examination. The 3-D distortion array pattern is plotted. Examination for Yagi-Uda radio antenna is executed and is extended to the advancement of aperture thinned radiation pattern of the Indian MST radar gathering mechanical assembly display. Assortment of parameters with different array thinning configuration is classified in Table 1. The scope of the work can be extended to polar plots.

Acknowledgments I particularly thank MST radar center facility for permitting me to complete the work. I like to astoundingly thank my guide and Scientist—MST radar center, Dr. P. Srinivasulu and Prof. Dr. N.C. Eswar Reddy, SVUCE, Tirupathi for helping me to complete the work viably in scheduled time.

References

1. P Srinivasulu: Handbook on MST Radar Center Facility.
2. Thiele G A: Analysis of Yagi-Uda Type Antennas, IEEE Trans on AP, Vol AP-17, 1969.
3. Karthik Kumar R, Sudha A, Gunasekaran N: Design and Analysis of Modified Taylor Distribution Adaptive Beamforming Array Element, IEEE International Conference on Microwave and millimeter Wave Technology, pp. 1–4 (2007).
4. N Dinesh Kumar: Principal Plane Pattern Distortion Analysis due to Array Thinning, International Journal of Applied Engineering Research, Vol 9, No. 19, pp. 4359–4368 (2014).
5. N Dinesh Kumar: Array Factor Distortion Analysis, Global Journal of Advanced Engineering Technology, Vol. 3, Issue 2, pp. 156–159 (2014).
6. Balanis C A, Antenna Theory Analysis & Design, 2/3. Jon Wiley & Sons.
7. Richmond J, Digital Computer solutions of the rigorous equations for scattering problems, Proc IEEE, Vol 53, pp. 796–804, Aug 1965.

Elasticsearch and Carrot^2-Based Log Analytics and Management

Prashant Kumar Singh, Amol Suryawanshi, Sidhharth Gupta and Prasad Saindane

Abstract The log analytics and management mechanism is an implementation of Elasticsearch—a distributed full-text search engine for indexing the logs and Carrot^2 clustering engine for clustering the logs which are combined together with our algorithm to manage and analyze the logs given to it as input. In this paper, we reflect on how Elasticsearch along with Carrot^2 is used with our algorithm to manage and analyze logs of any format. The log analytics and management is set up on Amazon web server.

Keywords Amazon Web Services · Elasticsearch · Carrot^2 · Data-Driven Documentation · Logs · Log analytics · Cloud · Clustering · Real-time Search · Real-time Analytics

1 Introduction

This paper provides an overview of Elasticsearch, Carrot^2, and real-time log analytics and management system. This system is being developed in order to implement a log analytics and management engine.

To study and derive information from large amounts of data is the most important analytics in today's scenario of machine generated data. This system will index, cluster the data and interpret it to perform policy-based analysis. The system

P.K. Singh (✉) · A. Suryawanshi · S. Gupta · P. Saindane
Department of Computer Engineering, MIT, Pune, Maharashtra, India
e-mail: prashantprs93@gmail.com

A. Suryawanshi
e-mail: amols203@gmail.com

S. Gupta
e-mail: siddharthgupta0409@gmail.com

P. Saindane
e-mail: saindaneprasad@gmail.com

H.S. Saini et al. (eds.), *Innovations in Computer Science and Engineering*,
Advances in Intelligent Systems and Computing 413,
DOI 10.1007/978-981-10-0419-3_9

will formulate the data for ready reference, making smart decisions, and will be helpful to understand the vast potential of cloud computing and Internet applications. With the growth of Internet, such type of systems are almost necessary for every cloud-related application which will give the maximum technical and business analytics benefits in future.

In a cloud computing model, n number of virtual servers are running and your application or website gets hosted by these servers. Logs get generated by every request to the servers, either virtual or real. These logs contain information about client, its request type, and IP address along with timestamp.

Let us take an example with logs generated by AWS. In the case of AWS, logs get generated for every action on AWS console like logging into console, accessing any service, changing attributes of any running service, etc.

For EC2 service of AWS, logs get generated when client performs request activities on the server. Various types of logs such as access logs, error logs, critical logs, and delete logs get generated. All these logs are indicators of the behavior of the system and the servers. Similar to these logs, millions of logs are generated on a daily basis by the systems and servers present everywhere. All these data contain valuable information in some way or the other.

2 Technologies and Tools

The different technologies and methods adopted to implement the log analysis for Amazon Web Services are given below.

2.1 Amazon Web Services

2.1.1 Amazon Elastic Cloud Compute (EC2)

Amazon EC2 [1] enhances the scalability by providing a scalable computing capacity in AWS cloud. It prevents the use of hardware and provides virtual scalable servers, configures security, and manages storage and networking.

2.1.2 Amazon Elastic Load Balancer (ELB)

Elastic load balancer [2] is used to distribute the incoming traffic across various EC2 instances. It also provides adding and removing the EC2 instances as the requirement changes. It can also encrypt and decrypt, thus enabling the servers to focus on their main task.

2.1.3 Amazon Simple Storage Service (S3)

Amazon S3 [3] is the object storage module of Amazon Web Services. It provides storage via web services interfaces, scalability and ubiquitous access to data, and high availability and low latency at low costs.

2.1.4 Amazon CloudTrail

AWS CloudTrail [4] tracks and records any activity which we perform in our AWS account. It delivers the records in the form of log files which are actually the AWS API calls we made in our account. It includes the identity of the API caller, the time of the API call, the source IP address of the API caller, the request parameters, and the response elements returned by the AWS service. It just provides the history of your AWS account.

2.2 PuTTY

PuTTY is nothing but a free Telnet and SSH terminal software for Windows and UNIX platforms that enables users to remotely access computers over the Internet.

2.3 Elasticsearch

Elasticsearch [5] is a querying tool and database server, written in Java. Elasticsearch is based on Apache Lucene and uses it for almost all of its operations. Elasticsearch is reliable, scalable, and multitenant. We use Elasticsearch as a backend programming tool and Querying our database. The logs accessed are indexed using elasticsearch. It converts the log in the inverted index format. This helps in writing different queries for log analysis. These queries give detailed results about the logs according to the specified requirements provided in the query.

2.4 Carrot^2

Carrot^2 [6] is an open source clustering tool. It clusters similar documents and gives them human readable labels. Carrot provides different algorithms for clustering like K-means, Lingo, and STC. Carrot^2 automatically clusters small collections of documents, e.g., search results or document abstracts into thematic categories. Here, Carrot^2 clusters the data provided by the elasticsearch queries specified. Apart from this, Carrot^2 specifies the clustering algorithm used to cluster the data and provides visual output for the same.

2.5 Data-Driven Documents (D3)

D3.js [7] is a JavaScript library which performs the binding between user data and Document Object Model (DOM). It is used for graphical representation of user data. It also allows the user to manipulate the data.

2.6 AWS CLI

The AWS Command Line Interface (CLI) is a tool provided by Amazon to unify all of its modules in AWS. With the help of Amazon CLI, a user can control multiple modules of AWS via terminal.

2.7 CRON

CRON is a time-based job scheduler specifically designed for UNIS like operating systems. It can be used to schedule all kinds of scripts, processes, and applications. It is commonly used for system automation and administration. We use CRON for periodically downloading logs from S3 bucket to our EC2 instance for indexing and clustering.

2.8 EC2 Instance

Elastic Compute Cloud (EC2) [1] instance is used to implement the functionalities of different tools and methods. It is accessed using an SSH and telnet client PuTTY which works as a command line interpreter for application of EC2 instance on a local machine.

3 Working

The logs are stored in Simple Storage Service of AWS. The logs are fetched using AWS CLI into the EC2 instance. If the fetched logs are in raw format they are converted to JSON format using our algorithm; if they are already in JSON format then the further processing is done.

The analytics and management mechanism works in three steps:

1. Conversion of logs into proper format using shell script.
2. Indexing of logs for faster retrieval.
3. Graphical representation.

Elasticsearch is a distributed full-text search engine that is used for indexing the logs. Indexing is done for faster retrieval of data. The Carrot^2 clustering engine is used along with Elasticsearch. Elasticsearch indexes the logs and the logs are clustered after a search query is fired. Document is the basic unit of Carrot^2 [8] clustering engine. Elasticsearch considers one log as one document and indexes it accordingly. The Carrot plugin groups together similar "documents" and assign labels to these groups. Each document consists of logical parts. For clustering document, the identifiers are specified by us according to the user demands.

As the indexing and clustering mechanism are integrated into one unit, the process of log analytics becomes really fast. The clustered data is then processed with our algorithms.

A graphical representation is done by using Data-Driven Documentation. A JavaScript library dc.js is used for this purpose. The processed result is integrated with dc.js and the parameters are defined for graphical plotting of the data, which is all done according to the user needs.

This kind of flexibility and faster processing power makes this log analytics engine unique.

3.1 Strengths

3.1.1 Speed

The speed of processing the logs is quite fast as the logs are pre-indexed and the clustering and indexing mechanism are integrated in one unit.

3.1.2 Flexibility

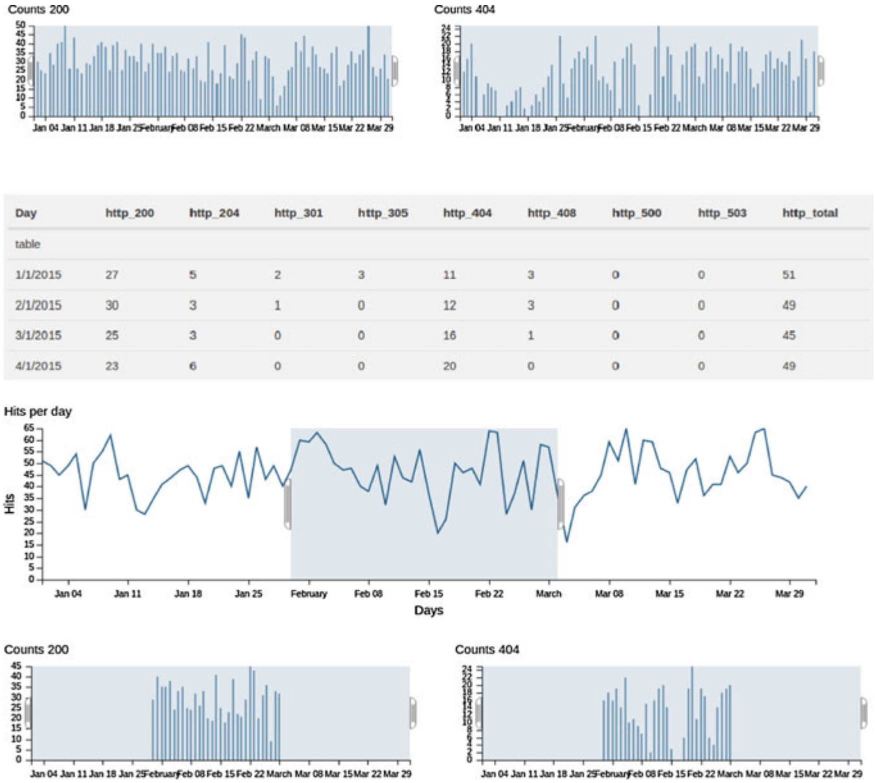
The log analysis and management is completely user-dependent. Every important aspect here can be customized by the user.

3.1.3 Scalability

As the log analytics and management engine are hosted on AWS, the server boasts of a great deal of scalability. The Elastic Compute Cloud ensures the engine is scalable in both the directions.

3.2 *Figures and Tables*

See Figs. 1 and 2.


Counts 200
50
45
40
35
30
25
20
15
10
5
0
Jan 04 Jan 11 Jan 18 Jan 25 February Feb 08 Feb 15 Feb 22 March Mar 08 Mar 15 Mar 22 Mar 29
Counts 404
24
22
20
18
16
14
12
10
8
6
4
2
0
Jan 04 Jan 11 Jan 18 Jan 25 February Feb 08 Feb 15 Feb 22 March Mar 08 Mar 15 Mar 22 Mar 29

Day	http_200	http_204	http_301	http_305	http_404	http_408	http_500	http_503	http_total
table									
1/1/2015	27	5	2	3	11	3	0	0	51
2/1/2015	30	3	1	0	12	3	0	0	49
3/1/2015	25	3	0	0	16	1	0	0	45
4/1/2015	23	6	0	0	20	0	0	0	49

Hits per day
65
60
55
50
45
40
35
30
25
20
15
10
5
0
Hits
Jan 04
Jan 11
Jan 18
Jan 25
February
Feb 08
Feb 15
Feb 22
March
Mar 08
Mar 15
Mar 22
Mar 29
Days
Counts 200
45
40
35
30
25
20
15
10
5
0
Jan 04 Jan 11 Jan 18 Jan 25 February Feb 08 Feb 15 Feb 22 March Mar 08 Mar 15 Mar 22 Mar 29
Counts 404
24
22
20
18
16
14
12
10
8
6
4
2
0
Jan 04 Jan 11 Jan 18 Jan 25 February Feb 08 Feb 15 Feb 22 March Mar 08 Mar 15 Mar 22 Mar 29

Fig. 1 Graphical representation of logs generated by AWS CloudTrail and elastic load balancer

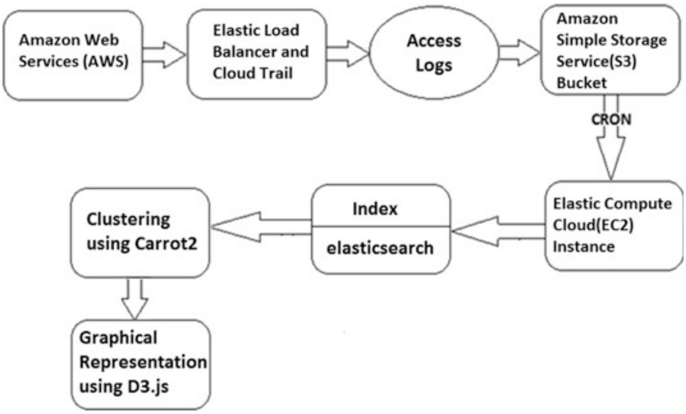

Amazon Web Services (AWS)
Elastic Load Balancer and Cloud Trail
Access Logs
Amazon Simple Storage Service(S3) Bucket
CRON
Elastic Compute Cloud(EC2) Instance
Index
elasticsearch
Clustering using Carrot2
Graphical Representation using D3.js

Fig. 2 Architecture of the log analytics and management engine

3.3 Program Code

3.3.1 Code to Index Logs

```
var sampledata;
doIndex = (function() {
var i = 0;
function doIndex(progressFn) {
var url = "/final/project/" + i;
var data=sampledata[i];
var current = i;
if (i < sampledata.length) {
i++;
$.post(url, JSON.stringify(data), function(result) {
doIndex(progressFn);
});
}
progressFn && progressFn(current, sampledata.length);
}
return doIndex;
})();

// ES search request data
var request = {
    "search_request": {
     "fields" : [ "timestamp", "elb", "backend_status_code","request"],
     "query" :{
             "filtered":{
                     "query" : {
                             "term" : {
                                             "elb" : "project"
                                     }
                             },
                     "filter" : {"range" : {"timestamp" : { "gte" : date1, "lte" :
date2}}}
                      }
            } ,
     "size": 1000 },
     "query_hint":"project",
     "algorithm": "stc",
     "field_mapping": { "title": ["fields.backend_status_code"] }
        };
```

3.4 Conclusion

The web server monitoring tool works in the phases of logging, parsing, and analyzing the logged data of the server. The web server monitoring tools provide a utility to monitor the web server with the real-time and time-based statistics of the server performance. The server statistics tracking, response code tracking, bandwidth statistics of the server, and CPU load statistics provide a detailed performance functioning of the server with present and historical data analysis of the logs.

3.5 Future Scope

(1) To study and derive information in large amounts of data is the most important analytics in today’s scenario for machine generated data.
(2) This engine will index and cluster the data and interpret it for policy-based analysis.
(3) The engine will formulate the data for ready reference, making smart decisions, and will be helpful to understand the vast potential of cloud computing and Internet applications.
(4) With the growth of Internet such type of engine is almost necessary for every cloud-related application which will give the maximum technical and business analytics benefits in future.

References

1. Amazon Elastic Cloud Compute Documentation, http://aws.amazon.com/ec2.
2. Amazon Elastic Load Balancer Documentation, http://aws.amazon.com/elb.
3. Amazon Simple Storage Service Documentation, http://aws.amazon.com/s3.
4. Amazon Cloudtrail Documentation, http://aws.amazon.com/cloudtrail.
5. ElasticSearch Documentation, http://www.elastic.co/guide.
6. Carrot^2 Documentation http://project.carrot2.org/documentation.html.
7. D3 Documentation, d3js.org.
8. GitHub Documentation for Carrot^2 Plugin, https://github.com/carrot2/elasticsearch-carrot2/.

High-Security Pipelined Elastic Substitution Box with Embedded Permutation Facility

K.B. Jithendra and T.K. Shahana

Abstract In today's world a major percentage of information and resource sharing is carried out through electronic means. A good portion of the above said activities needs security essentially. Since all the resources for any technology is provided by the nature, directly or indirectly, it is the moral responsibility entrusted with researchers to bring out maximum efficiency with minimal resources. In this paper a novel method is proposed in which Substitution Box is operated in a pipelined fashion, which can optimize hardware complexity and speed of the Block Ciphers. Facility for key-based permutation of message bits is integrated with the design, which offers additional security. The complexity and security analysis is also done.

Keywords Block cipher · S Box · Hardware complexity · Security · Diffusion · Pipeline · Elasticity

1 Introduction

Block ciphers are used extensively in encryption process in which the substitution box (S Box) is an indispensible part. S Box is the most important element in a block cipher because its contribution to security through its nonlinear nature. The basic purpose of an S Box is to introduce confusion. An S Box can be designed in many fashions, but unless and until its properties are consistent with certain criteria, which is discussed in the forthcoming sessions, it would not be effective. In block ciphers like AES or DES, parallel S Boxes are used which demands a greater level

K.B. Jithendra (✉)
College of Engineering, Vadakara, Calicut (Dt), Kerala, India
e-mail: jithendrakb@yahoo.com

T.K. Shahana
School of Engineering, Cochin University of Science and Technology,
Cochin, Kerala, India
e-mail: shahanatk@cusat.ac.in

H.S. Saini et al. (eds.), *Innovations in Computer Science and Engineering*,
Advances in Intelligent Systems and Computing 413,
DOI 10.1007/978-981-10-0419-3_10

of hardware complexity. Though serial processing of data with S Box is also possible, this reduces the speed of processing to a great extend. In this paper, the design is carried out based on the concept parallel processing and pipelining, which can bring an optimum operation point in which the trade-off between hardware complexity and speed of operation can be utilized effectively.

The additional feature introduced in this design is the permutation of message bits. The input data (message) is pipelined and data in each pipeline is serialized in a key-based random fashion. This really enhances the uncertainty of the block cipher. The permutation of the message bits does not require much hardware in addition to what is required for serialization of the bits.

This paper is organized as follows. Section 2 gives an overview about the conventional S Box and properties. Section 3 introduces the proposed design and explains how it functions. Experimental results, analysis and comparison with existing system are given in Sect. 4. Section 5 concludes the topic with future scope.

2 Conventional S Box and Properties

Encryption by mere substitution was existed even in ancient times. But later a scientific background was given to the same by Shannon [1]. All block ciphers use certain kind of S Boxes. The design of an S Box is crucial because a significant level of security is contributed to the block cipher is through the S Box. An S Box is supposed to satisfy a majority of the following properties.

- Nonlinearity: Any linear relations existing between input and output vectors ease the process of finding out the mapping. Nonlinearity is one of the most important properties which contribute greatly to security.
- Bijection: For an $n \times n$ S Box, Bijection means each data is substituted by a unique data [2].
- Balance: The function is a balanced one, when the truth table of the function carries equal number of zeros and ones. An S Box $S : \{0,1\}^n \rightarrow \{0,1\}^m$ is said to be balanced if and only if when all the m output columns are balanced.
- Bit independent criteria or correlation immunity: This refers to the state in which there is no statistical dependency exists between output bits of output vector [2, 3].
- Completeness: A Boolean function $f : \{0,1\}^m \rightarrow \{0,1\}$ is complete when each output bits are a function of all input bits.
- Strict avalanche criteria (SAC): Webster and Tavares [4] introduced the strict avalanche criteria (SAC). For introducing maximum confusion, a slight change in the input vector should be reflected as a significant change in the output vector. Whenever a single input bit is complemented at least 50 % bits of the output vector should change for satisfying SAC.

- Extended Properties—Static and Dynamic: Using information theory and previous works, Dawson and Tavares [5] proposed an extended set of desirable properties of S Boxes. Sivabalan, Tavares, and Peppard [6] proposed their own extended criteria. Here design criteria are proposed at multiple bits levels.

3 Proposed System: Pipelined Elastic S Box

The main aim of the proposed system is to reduce the hardware complexity and to enhance the uncertainty of block ciphers. This is achieved by introducing three new concepts which are not familiar in the conventional S Box design.

- Pipelining: The whole process is done in a pipelined fashion [7]. Each pipelines works serially. This technique effectively reduces the hardware complexity compared to parallel S box [8] and improves the speed of operation compared to serial S Box [9].
- Diffusion: The data is permuted among itself before substitution, based on the random key. Since data sampling is done in group of bits (hereafter mentioned as message block) random sampling will not cost much hardware.
- Elasticity: The size of the data bits taken for substitution varies randomly. The difference in size of message block of maximum length and minimum length is called elasticity. This concept considerably enhances the uncertainty.

Figure 1 shows the block diagram of the proposed system. The entire message is divided and fed to k number of pipelines (Message 1 to Message k) and processed in parallel. Each message group is processed serially to reduce the hardware complexity. Each message group is divided into different blocks of uneven size. The distribution of the block size can be varied in different message groups. The difference in block size between the largest block and the smallest block is called maximum elasticity e. If n is the block size of the largest block, the range of block sizes is from n to $n - e$. The controller takes care of the timing and control of different activities. For different block size of data, S Boxes with same input/output size is used. So the number of S Boxes used is $e + 1$, with size varying from n to $n - e$. In each message groups (pipelines), message blocks are selected serially in random fashion to introduce permutation of the bits and fed to substitution box. k number of random number generators are placed in the controller to provide unique pattern of message block selection for each pipeline. After substitution, the message blocks are appended in a serial fashion. Finally, the message groups are appended.

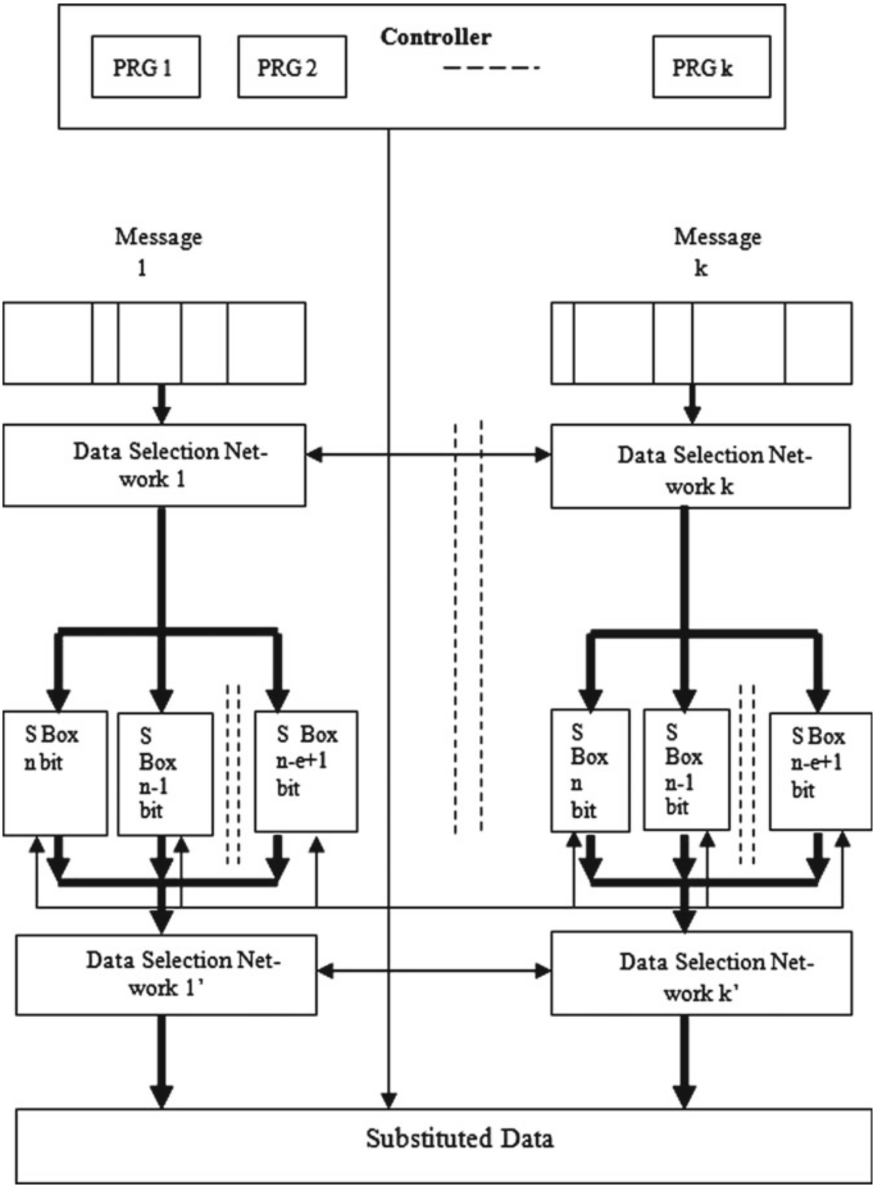


Fig. 1 Block diagram of pipelined elastic S Box

4 Experimental Results and Analysis

A 128 bit system in the proposed fashion is designed and implemented using Xilinx Design Suite 14.5 and Spartan 3A FPGA.

4.1 Data Analysis

Two pipelines are incorporated each having 15 message blocks of uneven size. An input data of 64 bits is applied to one path. The data diffused and data substituted for the same input with different keys are given in Table 1. D represents diffusion and S represents substitution. The table shows how the randomness is improved with the introduction of key-based permutation. The input message given in all cases is FEDC BA98 7654 3210 (Hex).

Table 1 Diffused and substituted data based on random key

Key	Process	Diffused/substituted data			
0001	D	2306	60EE	7AB7	FB90
	S	EF8B	1F6F	DE83	6338
0010	D	660D	D4AB	7FB9	0220
	S	B1FC	FDE8	3633	8EF8
0011	D	8118	3306	EA66	BFDC
	S	C77C	58FE	7EF4	1B19
0100	D	8330	6EA6	6BFD	C811
	S	C58F	E7EF	41B1	9C77
0101	D	55BF	DC81	1833	06EA
	S	F41B	1967	7C58	FEFE
0110	D	BA95	6FF7	2046	0CC1
	S	9FBD	06C6	71DF	163F
0111	D	9023	0661	BA96	6FFB
	S	38EF	8B1F	CFDE	8363
1000	D	1833	06EA	66BF	DC81
	S	7C58	FE7E	F41B	19C7
1001	D	C1BA	956F	F720	4606
	S	3F9F	BD06	C671	DF16
1010	D	B7FB	9023	0660	DD4A
	S	38,363	38EF	8B1F	CFDE
1011	D	52AD	FEE4	08C1	9837
	S	F7A0	D8CE	3BE2	C7F3
1100	D	3752	ADFE	E408	C198
	S	F3F7	A0D8	CE3B	E2C7
1101	D	FF72	0460	CC1B	A956
	S	6C67	1DF1	63F9	FBD0
1110	D	FB90	2306	605D	4AB7
	S	6338	EF8B	1FCF	DE83
1111	D	DC81	1833	06EA	55BF
	S	19C7	7C58	FE7E	F41B

4.2 Cryptanalysis

Linear Cryptanalysis This method checks the existence of any linear relations between input and output vectors. Each pipeline uses S Box's of variable size regardless the data length of pipeline. The linear cryptanalysis faces 2 additional difficulties with the proposed design.

- The probability biases of each S Box will be different which makes the calculations using Piling up principle [10] difficult.
- In conventional design every bits are processed simultaneously. Differing from the conventional SPN, message blocks in the proposed design are in queue. The S Box's here are 'reused' for data in a single pipeline. Hence establishing the relationships of data in consecutive rounds is very difficult.

Differential Cryptanalysis Differential cryptanalysis checks for any relations existing between the differences of consecutive input vectors and that of output vectors. For this to carry out.

- The differential distribution table of different S Box's will be different
- Here also the reuse of S Box's makes it difficult to establish relations between consecutive rounds.

4.3 Timing Issues

Since each pipeline is working serially, the proposed design is slower than that of conventional systems. Number of message blocks determines the time with which each pipeline completes the process. The above-implemented design takes 15 clock cycles for the diffusion and substitution of 64 bit massage. Increasing number of pipelines speeds up the process but increases the hardware complexity as well.

4.4 Comparison with Existing Systems

The hardware complexity of the proposed design is given in Table 2. Hardware complexity of an AES S Box is given in Table 3. Both represent 128 bit systems. From the tables, it is clearly evident that the proposed system consumes less hardware.

Table 2 Device utilization summary of proposed system —Spartan 3A

Device utilization summary			
Logic utilization	Used	Available	Utilization (%)
Number of slices	100	5888	1
Number of slice flip flops	26	11,776	0
Number of 4 input LUTs	188	11,776	1
Number of bonded IOBs	256	372	68
Number of GCLKs	1	24	4

Table 3 Device utilization summary of AES S Box—Spartan 3A

Device utilization summary			
Logic utilization	Used	Available	Utilization (%)
Number of Slices	1098	5888	18
Number of 4 input LUTs	2048	11,776	17
Number of bonded IOBs	258	372	69
Number of GCLKs	1	24	4

5 Conclusion and Future Scope

A novel concept of pipelined elastic Substitution Box is introduced in this paper with implementation details. High security as well as low hardware complexity is claimed for the proposed design. An additional feature of data permutation is also introduced without much additional hardware requirement. The concept of elasticity and diffusion effectively resists both differential and linear cryptanalysis. The hardware complexity of the proposed system is significantly less than that of conventional systems. Extended research is possible on this topic especially in integrating other cryptographic functions. Hash function generation also can be done effectively with the proposed design.

References

1. C. Shannon, Communication theory of secrecy systems, Bell Systems Technical Journal, vol. 28, 1949.
2. C Adams and S Tavares, The structured design of good S Boxes, Journal of Cryptology, 3 (1):27–41,1990.
3. J.Cobas and J.Brugos. Complexity theoretical approaches to the design and analysis of cryptographical Boolean functions In Computer Aided Systems Theory-EUROCAST 2005, LNCS. Springer-Verlag, Berlin, Germany, 2005.
4. A. Webster, S. Tavares, On the Design of S Boxes, Advances in Cryptology-CRYPT0 1985, LNCS 218, Springer-Verlag, 1985.

5. M. Dawson, S. Tavares, An Expanded Set of S Box Design Criteria Based on Information Theory and its Relation to Differential-like Attacks, Advances in Cryptology—EUROCRYPT 1991, LNCS 547, Springer-Verlag 1991.
6. M Sivabalan, S. Tavares, L. Peppard, On the design of SP networks from an information theoretic point of view, Advances in Cryptology—CRYPTO 1992, LNCS 740, Springer Verlag 1993.
7. Kai Hwang and Faye A. Briggs, Computer Architecture and Parallel processing, Tata McGraw-Hill, New Delhi 2012.
8. W. Stallings, Cryptography and Network Security, Principles and Practices, Prentice Hall, 2006.
9. Jithendra K.B, Shahana T.K, High Security Elastic Serial Substitution Box for Block Ci-phers, Proceedings of Second International Conference on Networking, Information and Communication (ICNIC), Bangalore 2015.
10. C K Shyamala, N Harini, T R Padmanabhan, Cryptography and Security, Wiley India Pvt ltd, New Delhi, 2011.

An Integrated Approach to High-Dimensional Data Clustering

Rashmi Paithankar and Bharat Tidke

Abstract Applying the traditional clustering algorithms on high-dimensional datasets scales down in the efficiency and effectiveness of the output clusters. H-K Means is advancement over the problems caused in K-means algorithm such as randomness and apriority in the primary centers for K-means, still it could not clear away the problems as dimensional disaster which is due to the high-computational complexity and also the poor quality of clusters. Subspace and ensemble clustering algorithms enhance the execution of clustering high-dimensional dataset from distinctive angles in diverse degree, still in a solitary viewpoint. The proposed model conquers the limitations of traditional H-K means clustering algorithm and provides an algorithm that automatically improves the performance of output clusters, by merging the subspace clustering algorithm (ORCLUS) and ensemble clustering algorithm with the H-K Means algorithm that partitions and merge the clusters based on the number of dimensions. Proposed model is evaluated for various real datasets.

Keywords H-K clustering · Ensemble · Subspace · High dimensional · Clustering

1 Introduction

As a prominent strategy in data mining, Clustering [1] is a preprocessing step for preparing data for subsequent processing. Clustering is utilized in applications such as data mining, pattern recognition, image processing, and machine learning [2, 3]. For high-dimensional data clustering, the well-known problem is dimensionality disaster occurs due to the problems such as increasing sparsity of data and increasing

R. Paithankar (✉) · B. Tidke
Department of Computer Engineering, Flora Institute of Technology,
Pune, Maharashtra, India
e-mail: paithankar.rashmi@gmail.com

B. Tidke
e-mail: batidke@gmail.com

H.S. Saini et al. (eds.), *Innovations in Computer Science and Engineering*,
Advances in Intelligent Systems and Computing 413,
DOI 10.1007/978-981-10-0419-3_11

difficulty in distinguishing distances between data points makes clustering difficult. The proposed model combines the advanced techniques as subspace clustering and ensemble clustering algorithm with traditional H-K Means clustering algorithm and their advantages to improve the performance of clustering result on high-dimensional data. The proposed model uses ORCLUS—subspace clustering algorithm. H-K means algorithm is applied on each subspace that will generate the hierarchy of clusters. At each level of hierarchy there will be comparison of the MSE of parent and child cluster, based on the comparison merge and ensemble stage is applied. This comparison based merge and ensemble of clusters will provide the consistency in the size and accuracy of clusters. And finally, we will get the set of output clusters.

2 Related Work

Dimension reduction, subspace clustering, ensemble clustering and K-means clustering [4–6] are some of the areas of clustering. Dimension reduction has the problems such as loss of most of the information and could not produce effective clusters when data is in different subspaces. Subspace clustering is introduced to overcome the problem of data identifying in different subspaces which gives the solution algorithms as CLIQUE [7], PROCLUS [8], and MAFIA [9]. H-K (Hierarchical K-means clustering algorithm) clustering algorithm is introduced by Tung-Shou Chen et al. [10] overcomes the problems of traditional K-means clustering algorithm by combining the hierarchical clustering method and partition clustering method organically for data clustering. This provides improved clustering method which overcomes the problems as randomness and apriority of initial centers selection still the execution process requires high-computation complexity and the poor accuracy in output clusters. Ensemble clustering combines the solutions of many clustering algorithms based on the consensus function and achieve more pertinent solution.

3 Proposed Model

The high-dimensional datasets are given as the input to the proposed system; it will take the data through three steps which apply the advanced clustering algorithms.

3.1 Dataset Preprocessing

The real-time dataset—X, number of subspaces—k, and number of dimensions—l, is given as input to the proposed model. Replace the missing values in dataset and get the preprocessed data as output to process in the next steps. (Mostly prefer the real-time datasets over synthetic datasets for more accurate results).

3.2 The Subspace Clustering Process

Apply the subspace clustering algorithm—ORCLUS, on the preprocessed dataset X. First, the assignment stage assigns the data points to the preselected set of seeds S_i, and creates the set of clusters C_i. $\mathcal{E}_i$ is the set of vectors defines the q dimensional subspace for the cluster C_i. This subspace is chosen such that for that the cluster C_i should have the least spread. So, $\mathcal{E}_i$ subsist of the eigenvectors of the $l_0 \times l_0$ covariance matrix Σ_i of the points in C_i that correspond to the q smallest eigen values of Σ_i. $E(C_i, \mathcal{E}_i)$ is the projected energy of the cluster C_i in subspace $\mathcal{E}_i$ which is the sum of the eigen values [11]. The values α and β are the user defined factors which specifies the reduction in the number of clusters and the dimensionality of the subspace. The values of α and β are related by the following equation:

$$(\ln(m/m_0))/(\ln(l/l_0)) = \ln\alpha/\ln\beta. \tag{1}$$

```
Input: Input dataset ($X$), Number of subspaces ($k$), Number of dimensions ($l$).
Output: $k$ set of subspace
Steps:
1. Start with input dataset $X$
2. Pick a set of $m_0 > m$ points from the database $X$ and denote them by $S_0$.
3. For current cluster, Set $m_c = m_0$, $l_c = l_0$, $S_c = S_0$
4. Set $E_\mathrm{i}$, the subspace for current cluster $C_i$ as the set of vectors defining the initial feature space, for $i = 1,\ldots, m_c$.
5. Set $\alpha = 0.5$, and calculate the value of $\beta$ from Eq. (1).
6. While $m_c > m$ do
    • For each $i$, $i = 1,\ldots, m_c$, all the points in $X$ that are closer to $i$th element of $S_c$ are assigned to the cluster $C_i$. (The distance between two points is figured out in the $\mathcal{E}_i$ subspace.)
    • For each $i$, $i = 1,\ldots, m_c$ define $\mathcal{E}_i$ as the set of eigenvectors corresponding to the $l_c$ smallest eigen values of the $l_0 \times l_0$ covariance matrix of $C_i$.
    • Set $m_\mathrm{new} = \max\{m, \alpha m_c\}$ and $l_\mathrm{new} = \max\{l, \beta l_c\}$
    • For each pair $i$ and $j$ such that $i < j$; $i, j = 1,\ldots, m_c$,
      Determine the $\mathcal{E}_{ij}$ for the $C_i \cup C_j$ and $E(C_i \cup C_j, \mathcal{E}_{ij})$.
    • While $m_c > m_\mathrm{new}$ do
        - Determine $E(C_x \cup C_y, \mathcal{E}_{xy}) = \min i, j\text{–}1,\ldots, m_c, i \neq j\ E(C_i \cup C_j, \mathcal{E}_{ij})$ and merge $C_x$ and $C_y$ to $C_r = C_x \cup C_y$.
        - Recalculate the necessary $E(C_i \cup C_r, \mathcal{E}_{ir})s$ in light of the previous merging.
        - $m_c = m_c - 1$
        - End(while)
7. $m_c = m_\mathrm{new}$
8. Set $S_c$ as the means of the $m_\mathrm{new}$ clusters formed by the previous steps (While loop)
9. End (While)
```

3.3 The H-K Clustering Process and Splitting Process

The above stage will output the *S* set of clusters, apply the H-K Means clustering algorithm on them. Approach preferred by proposed model is the divisive H-K clustering algorithm.

Input: *k* set of subspace.
Output: *N* set of clusters.
Steps:
1. Process each subspace, start with S_1.
2. Calculate Mean *M* of S_1.
3. Divide the S_1 subspace into two Ranges: R_1 and R_2, R_1 contains the values lower than *M* and R_2 contains the values higher than *M*.
4. Apply K-Means Algorithm to each R_i ($i = 1, 2$).
5. Calculate the MSE of each R_i.

$$\mathrm{SSE} = \sum_{i=1}^{N} \sum_{x_i \text{€} c_i} \left(\left\| x_i - \mu_i \right\| \right)^2$$

Where, μ_i is the mean of cluster R_i and *x* is the data object belongs to R_i cluster. Formula to compute μ_i is shown in following equation,

$$\mu_i = \left(\frac{1}{n_i} \right) \sum x_i \text{€} R_i x_i$$

6. Process first subspace, R_1.
7. Repeat the steps (2) to (5).
8. Compare the MSE of parent and child, if the MSE of child is greater than parent go to step (7) else jump to step (9).
9. Process second subspace, R_2.
10. Repeat the steps (2) to (5).
11. Compare the MSE of parent and child, if the MSE of child is greater than parent go to step (10) else jump to step (12).
12. Go to step (1).
13. This set will output N set of clusters

3.4 Ensemble Clustering Process

For all the splitted clusters which are the output of the above step check again, the distance between them and if some clusters are near each other merge them based on the objective function (distance function). Also check mean square error (MSE) of each merged cluster with the parent cluster if found to be larger, that cluster must be unmerged and available to be merge with some other cluster in the hierarchy.

Input: *N* clusters.
Output: Partition $C_1,\ldots,C_c$
Steps:
1. Start with n node cluster.
2. Find the closest two cluster using Euclidean distance from the hierarchy and merge them.
3. Calculate MSE of root cluster and new merge cluster

$$\text{SSE} = \sum_{i=1}^{N}\sum_{x_i € c_i}\left(\left\|x_i - \mu_i\right\|\right)^2$$

where, μ_i is the mean of cluster R_i and x is the data object belongs to R_i cluster. Formula to compute μ_i is shown in following equation,

$$\mu_i = \left(\frac{1}{n_i}\right)\sum x_i € R_i x_i$$

4. In sum of squared error formula, the distance from the data object to its cluster centroid is squared and distances are minimized for each data object. Main objective of this formula is to generate compact and separate clusters as possible. If MSE of new merge cluster is smaller than the cluster after splitting keep it otherwise unmerges them. 5. Repeat until all possible clusters are merged according to step 4.

4 Implementation Details and Results

A clustering assessment requests solid measure for the appraisal and examination of clustering analyses and results. This section gives brief overview of methodology for processing, datasets which going to be used for experiments and the experimental results, and comparison of obtained experimental results for proposed model with the existing solutions. For file reading and ORCLUS execution, the concept of MULTITHREADING is used which increases the execution speed, comparative results for single threading and multi threading of the proposed model for wine dataset is shown in Fig. 2 and Table 2. For experimental evaluation, following datasets are considered. This datasets can be obtained from the UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/): 1. Wine dataset having 13 dimensions and 178 instances, 2. Yeast dataset containing 8 dimensions and 1484 instances (Fig. 1).

Tables 1 and 2 shows the variance of time and MSE values as per the input number of subspaces, dimensions, and according to the generated cluster numbers for wine and yeast dataset. Time of execution gets reduced for the proposed model by the usage of the split and merge strategy and the implementation of the MULTITHREADING technique for file reading and processing and the MSE values also get reduced (Table 2).

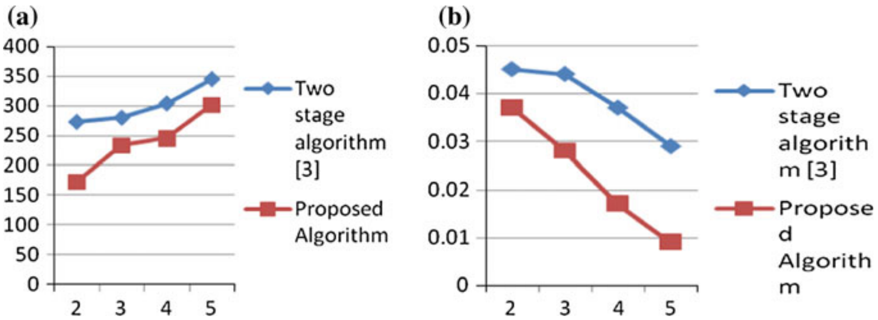


Fig. 1 **a** Comparison of time of execution of two-stage algorithm [3] with proposed algorithm for wine dataset. **b** Comparison of time of MSE of two stage algorithm [3] with proposed algorithm for wine dataset

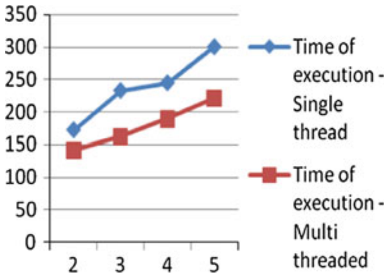


Fig. 2 Comparison of time of execution for single threading and multithreading for proposed algorithm on wine dataset

Table 1 Wine dataset results for, (a) two stage algorithm [3], (b) proposed model

Number of subspaces	Number of dimensions	Number of clusters	MSE values	Time of execution (MS)
(a) Two stage algorithm				
2	4	9	0.045	273
3	5	10	0.044	280
4	6	12	0.037	304
5	7	26	0.029	345
(b) Proposed model				
2	4	9	0.037	172
3	5	12	0.028	233
4	6	14	0.017	245
5	7	33	0.009	301

Table 2 Wine dataset results for proposed model (single threaded and multithreaded)

Number of subspaces	Number of dimensions	Number of clusters	Time of execution-single thread	Time of execution-multi threaded
2	4	9	172	141
3	5	12	233	162
4	6	14	245	189
5	7	33	301	221

5 Conclusion

High-dimensional data inherently incorporate the disadvantages as the curse of dimensionality and the sparsity of data. H-K Means accord good output clusters when the dataset is low dimensional but introduce high-computational complexity when applying the same algorithms on the high-dimensional dataset. Thus the proposed model provide the solution by associating the H-K Means algorithm with the subspace clustering algorithm that will find the output clusters from each subspace by considering each subspace as a different dataset which will simultaneously reduce the complexity as the search for cluster becomes in the small dataset with less dimensions. Applying the advanced clustering approach with the H-K Means will help to improve the performance of clustering process and will provide the stability of H-K Means clustering algorithm for high-dimensional data by improving the accuracy of output clusters.

References

1. McLachlan G., and Basford K., Mixture Models: Inference and Applications to Clustering, Marcel Dekker, New York, NY, 1988.
2. Shi Na, Li Xumin, Guan Yong Research on K-means clustering algorithm. Proc of Third International symposium on Intelligent Information Technology and Security Informatics, IEEE 2010.
3. Vance Faber, Clustering and the Continuous k-Means Algorithm, LosAlamos Science, 1994.
4. A.K. Jain, M.N. Murty, and P.J. Flynn 1999, -Data Clustering: A Review, ACM Computing Surveys, vol. 31, no. 3, pp. 264–323.
5. Zhizhou KONG et al. 2008, A Novel Clustering-Ensemble Approach, 978-1-4244-1748-3/08/ IEEE
6. Weiwei Zhuang et al. Ensemble 2012, Clustering for Internet Security Applications, IEEE transactions on systems, man, and cybernetics part c: applications and reviews, vol. 42, no. 6.
7. R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan 1998, Automatic subspace clustering of high dimensional data for data mining applications‖, In Proceedings of the 1998 ACM SIGMOD international conference on Management of data, pages 94–105. ACM Press.
8. B.A Tidke, R.G Mehta, D.P Rana 2012, A novel approach for high dimensional data clustering, ISSN: 2250–3676, [IJESAT] international journal of engineering science & advanced technology Volume-2, Issue-3, 645–651.

9. Guanhua Chen, Xiuli Ma et al. 2009, Mining Representative Subspace Clusters in High-Dimensional Data, Sixth International Conference on Fuzzy Systems and Knowledge Discovery.
10. Derek Greene et al. 2004, Ensemble Clustering in Medical Diagnostics, Proceedings of the 17th IEEE Symposium on Computer-Based Medical Systems (CBMS'04) 1063–7125/04.
11. K. A. Abdul Nazeer & M. P. Sebastian Improving the Accuracy and Efficiency of the K-Means Clustering Algorithm. Proceedings of the World Congress on Engineering 2009 Vol I WCE 2009, London, U.K, July 1–3, 2009.

Performance Analysis of Network Virtualization in Cloud Computing Infrastructures on OpenStack

Vo Nhan Van, Le Minh Chi, Nguyen Quoc Long and Dac-Nhuong Le

Abstract Cloud computing has become popular in IT technology because of advantages that focus on flexible, scaling, resources and services which help customers easy to build their own on-demand IT system. Cloud computing also has ability to balance, share, and manage IT resources between customers to get better performance. OpenStack, a new open source cloud computing framework which was a built-in modular architecture and focus on IaaS. OpenStack also focuses on NaaS by using network virtualization technology and OpenStack has been used popular in business. This paper does a research on network performance on OpenStack network module code name Neutron. The parameter related to network performance such as throughput, package loss, time and delay of data transmission are estimated through UDP protocol. Our research investigated the possible internal traffic flow pattern and evaluated network performance of each pattern on OpenStack cloud computing environment.

Keywords Cloud computing · Network virtualization · Network performance · Openstack

V.N. Van · N.Q. Long
Duytan University, Danang, Vietnam
e-mail: vovannhan@duytan.edu.vn

N.Q. Long
e-mail: nguyenquoclong@duytan.edu.vn

L.M. Chi
Danang ICT Infrastructure Development Center, Da Nang, Vietnam
e-mail: chilm@danang.gov.vn

D.-N. Le (✉)
Haiphong University, Haiphong, Vietnam
e-mail: Nhuongld@hus.edu.vn

H.S. Saini et al. (eds.), *Innovations in Computer Science and Engineering*,
Advances in Intelligent Systems and Computing 413,
DOI 10.1007/978-981-10-0419-3_12

1 Introduction

Cloud computing is a technology trend in recent years and known as a new approach in IT investment from infrastructure to software. Cloud computing delivers three main service models which are IaaS (*Infrastructure as a Service*), PaaS (*Platform as a Service*), SaaS (*Software as a Service*) and depend on each case of customers' requirements on storage, computing, software service, etc. [1, 2].

In IaaS cloud service model, virtualization technology is the core component which combines with a cloud management layer to satisfy requests from users. However, performance in IaaS model is very important to pursue moving from traditional environment to cloud. One of important concerns in IaaS model is network performance because the network function and network virtualization have responsibility for the whole cloud network operation, included customers' data transmission. Each cloud computing solution usually has separate network function to do this task. OpenStack developed a module named Neutron for network visualization in cloud [3].

The paper outlines as follows: the first section introduces basic information about cloud computing, the second section presents OpenStack—an open source cloud operating system, we focus on network performance analysis in open source cloud OpenStack in the third section, and finally is the conclusion and future works.

2 OpenStack

In this section, we focus on benefits of open source cloud deployment when compared with other commercial cloud. Before going to OpenStack solution, some benefits from open source cloud computing compared to commercial solution were analyzed in the documents from Canonical [4] or CSA [5]. The main features of open source cloud deployment are:

- *Avoiding vendor lock-in*: open source solution use open standards such as APIs, image format, special storage when compared to built-in features in commercial solution. This makes open source solution as high integrated with other system.
- *Community support*: interested open source solutions usually are supported by thousands of programmers in the world. This is one of the advantages that cannot appear in commercial solutions.
- *Scalability and licensing*: deploy and maintenance of an open source solution is more difficult, but it is always free and no charge for license. Moreover, open source solutions also do not require license for system scaling when compared with commercial solutions.
- *Modules development*: open source solution supports open standard for developing and integrating new software modules. In commercial solution, customer needs to order with high cost or waiting for a new upgrade from vendor to have new features.

2.1 OpenStack Architecture

OpenStack [6] is an open source cloud operating system for building private and public clouds. OpenStack is totally open source and designed to be a very large IaaS private cloud with large network and virtual machines. *The OpenStack mission: Provide a popular open source cloud computing framework in order to support all types of public or private cloud computing with variety system sizing but simple in deployment and high scalability* [4]. According to the mission above, OpenStack is a cloud computing controller which control IT resources such as compute, storage, and network in a data center. All of the operations on IT resources are delivered through a control panel which helps administrators control and interact with IT resources through a web based. OpenStack [6] has a 6-month period development, has new version, each 6 months. Each new version usually improves previous function for stability and deploy new functions. OpenStack is free totally and its components are written by Python language. The logical architecture of OpenStack is shown in Fig. 1.

In which five main components of OpenStack consist of Compute (*Nova*), Object Storage (*Swift*), Image Service (*Glance*), Dashboard (*Horizon*), and Identity (*Keystone*). OpenStack development is still in progress. OpenStack has full of cloud computing features that provide IaaS. Compute provide and manage compute function for instances (*Virtual machines*). Image Service store image of instance, is used by Nova when deploying an instance. Object Storage provide storage function. Dashboard provides web based for administration of OpenStack. Identity provide authentication and authorization for all of the services in OpenStack [6].

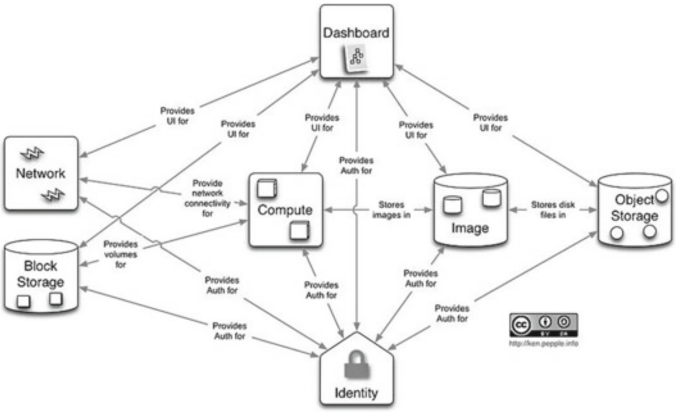


Fig. 1 OpenStack architecture

2.2 OpenStack Networking

OpenStack networking provide a variety of API to manage and define connections in cloud infrastructure. Neutron support many network technology to connect to cloud infrastructure. This part introduces Neutron, and basic services of Neutron in cloud infrastructure. Neutron includes three components which are Network, Subnet, Port.

- *Network*: manage network in Layer 2, the same as VLAN in physical network.
- *Subnet*: a group of IP v4/v6 which is configured for cloud system.
- *Port*: a port is defined and attached to a virtual device in cloud system.

End users can create their own network topology through Neutron. They create Network and subnet and attach between port and device in cloud system to connect their own system. Neutron allow end users create private network and attach to tenants, a tenant can have more than one private network and private network can be the same in each tenants. Some services of Neutron include:

- Create a network topology for end user system. For example: create web server system has many tier, for backup or load balancing.
- Provide flexibility in network management, allow end user edit and optimize their need.
- Provide extent API, allow programmers develop and integrate Neutron into other different system

Neutron support many network technologies to enhance network functions, not only just create VLAN, subnet, but also create virtual switch through vSwtich technology abd other features such as firewall, DHCP, VPN, load balancing.

In [7–9], the authors had deployed Hadoop on the OpenStack to compare the network performance of single-host plan and the multi-host plan through the service performance of Hadoop. In the latest paper, we had been analyzed performance of OpenStack open source solution for IaaS cloud computing in [10].

3 Network Performance Analysis of OpenStack

3.1 Our Topology Experiments

Open source cloud OpenStack support VPS service (IaaS) with different types of vCPU, RAM, and HDD. In our setup environment, the four testing cases will be analyzed to estimate the network performance in throughput, time delay, and packet loss in UPD protocol between virtual machines. Our deployment system has three servers which are:

- 01 server is Controller and Network roles, run services of OpenStack Controller management and shared services such as KeyStone, MySQL, Neutron, Cinder, and other related services.

- 02 servers are Compute nodes, which are deployed Neutron L2 Agent and the virtualization hypervisor KVM. These 02 servers support the compute service for virtual machines.

The deployment model has four networks:

- Public and Floating IP (*VM Network in deployment model*): 10.196.205.0/24
- Management network: 10.196.202.0/24
- Storage network: 10.196.203.0/24. In deployment model, the storage network is in the same physical interface (NIC) with Management Network. However, the storage network is not analyzed in this paper.
- Internal Network—network for tenants: 192.168.111.0/24 versus 192.168.112.0/24
- Hence, 1 PXE network is used for administration and installation of physical machine in the system.

3.2 Our Experiments

This section includes the four tests result found from the actual experimental setup on our testing deployment system in Fig. 2.

Case study 1. This experiment case does a testing in throughput, delay, and package loss between two virtual machines in the same compute node, the same internal network in UDP protocol.
Case study 2. This experiment case does a testing in throughput, delay, and package loss between two virtual machines in the same compute node but different internal network in UDP protocol.

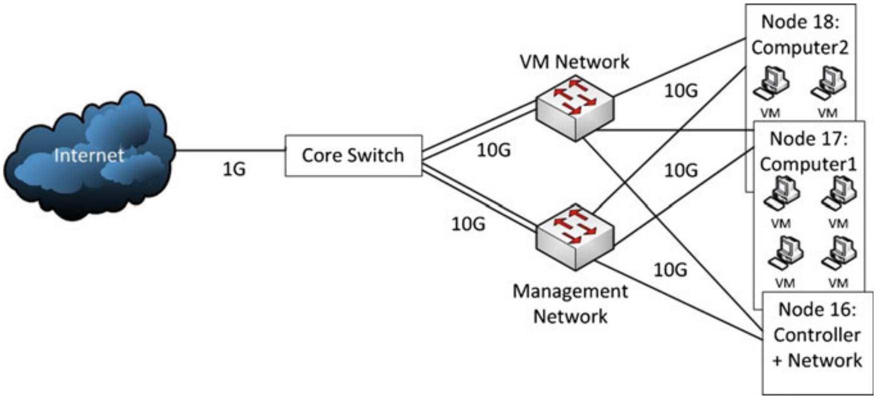
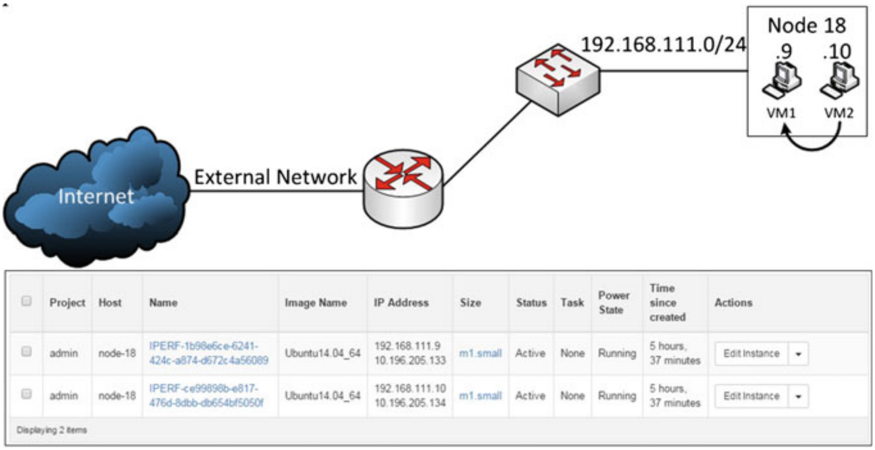


Fig. 2 Our testing deployment system

Case study 3. This experiment case does a testing in throughput, delay, and package loss between two virtual machines in different compute nodes but the same internal network in UDP protocol.

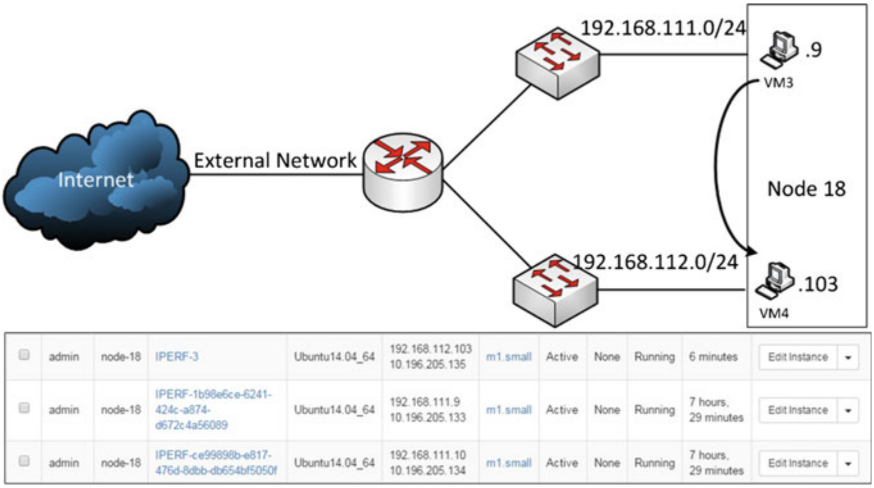
Case study 4. This experiment case does a testing in throughput, delay, and package loss between two virtual machines in different compute nodes and different internal networks in UDP protocol (Figs. 3, 4, 5, 6).



□	Project	Host	Name	Image Name	IP Address	Size	Status	Task	Power State	Time since created	Actions
□	admin	node-18	IPERF-1b98e6ce-6241-424c-a874-d672c4a56089	Ubuntu14.04_64	192.168.111.9 10.196.205.133	m1.small	Active	None	Running	5 hours, 37 minutes	Edit Instance ▾
□	admin	node-18	IPERF-ce99898b-e817-476d-8dbb-db654bf5050f	Ubuntu14.04_64	192.168.111.10 10.196.205.134	m1.small	Active	None	Running	5 hours, 37 minutes	Edit Instance ▾

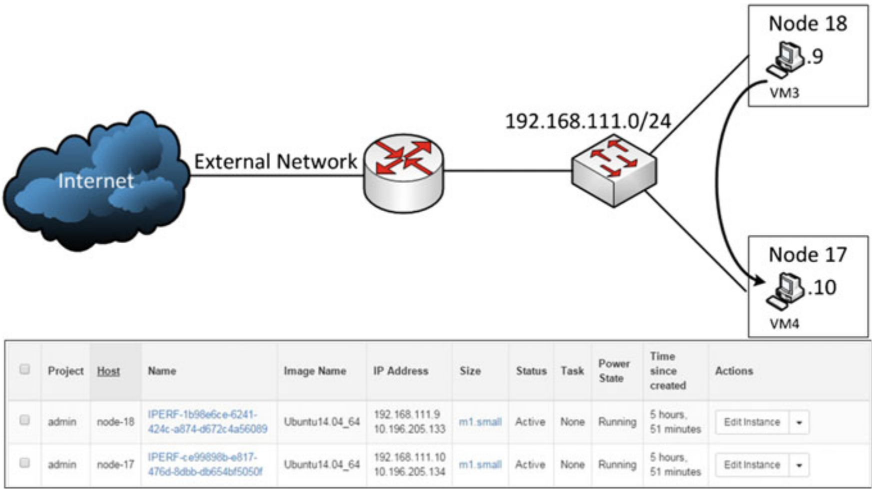
Displaying 2 items

Fig. 3 Case study 1: two virtual machines VM_1 and VM_2 in the same network 192.168.111.0/24 and compute node 18



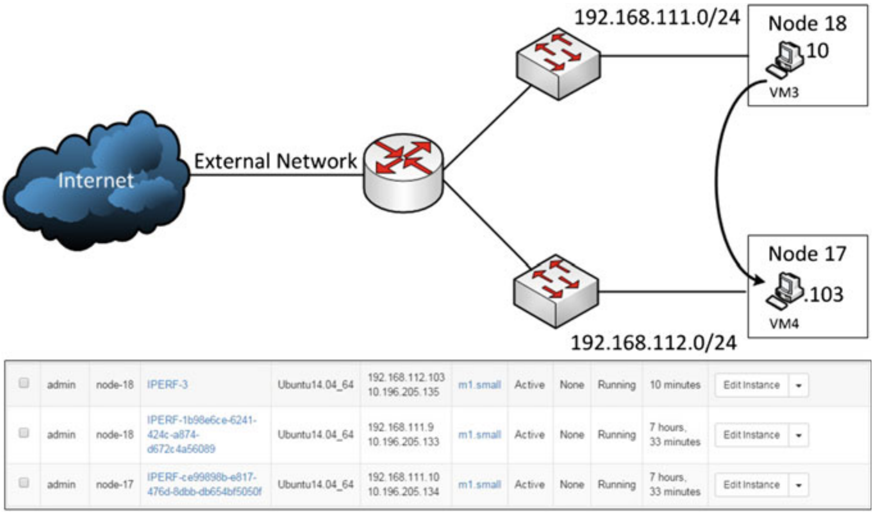
□	admin	node-18	IPERF-3	Ubuntu14.04_64	192.168.112.103 10.196.205.135	m1.small	Active	None	Running	6 minutes	Edit Instance ▾
□	admin	node-18	IPERF-1b98e6ce-6241-424c-a874-d672c4a56089	Ubuntu14.04_64	192.168.111.9 10.196.205.133	m1.small	Active	None	Running	7 hours, 29 minutes	Edit Instance ▾
□	admin	node-18	IPERF-ce99898b-e817-476d-8dbb-db654bf5050f	Ubuntu14.04_64	192.168.111.10 10.196.205.134	m1.small	Active	None	Running	7 hours, 29 minutes	Edit Instance ▾

Fig. 4 Case study 2: two virtual machines in the same node 18, but different internet subnet (192.168.111.9/24 and 192.168.112.103/24)



□	Project	Host	Name	Image Name	IP Address	Size	Status	Task	Power State	Time since created	Actions
□	admin	node-18	IPERF-1b98e6ce-6241-424c-a874-d672c4a56089	Ubuntu14.04_64	192.168.111.9 10.196.205.133	m1.small	Active	None	Running	5 hours, 51 minutes	Edit Instance ▾
□	admin	node-17	IPERF-ce99898b-e817-476d-8dbb-db654bf5050f	Ubuntu14.04_64	192.168.111.10 10.196.205.134	m1.small	Active	None	Running	5 hours, 51 minutes	Edit Instance ▾

Fig. 5 Case study 3: two virtual machines in the same internal network (192.168.111.0/24) but in 2 different compute nodes (node 17 and node 18)



□	admin	node-18	IPERF-3	Ubuntu14.04_64	192.168.112.103 10.196.205.135	m1.small	Active	None	Running	10 minutes	Edit Instance ▾
□	admin	node-18	IPERF-1b98e6ce-6241-424c-a874-d672c4a56089	Ubuntu14.04_64	192.168.111.9 10.196.205.133	m1.small	Active	None	Running	7 hours, 33 minutes	Edit Instance ▾
□	admin	node-17	IPERF-ce99898b-e817-476d-8dbb-db654bf5050f	Ubuntu14.04_64	192.168.111.10 10.196.205.134	m1.small	Active	None	Running	7 hours, 33 minutes	Edit Instance ▾

Fig. 6 Case study 4: two virtual machines in different compute nodes (node 17 and node 18) and different internal network (192.168.112.103 versus 192.168.111.10)

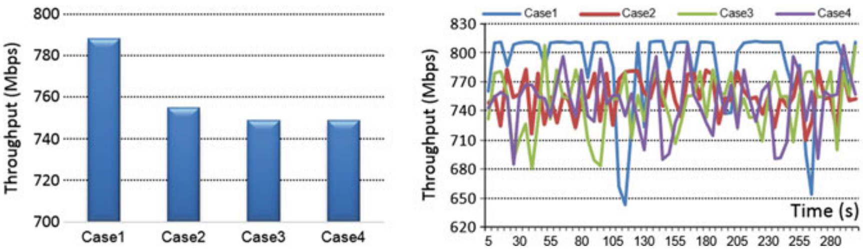


Fig. 7 Average UDP throughput in four experiment cases

3.3 Results and Discussion

Network function in OpenStack is delivered by Neutron module, and Neutron virtualize and create switching, routing environment through Neutron L2 Agent and Neutron L3 Agent service. All of the experiment cases above are designed and created by Neutron.

In research network parameter related to throughput, time delay, and package loss in UDP protocol, the tool IPERF is executed in 5 min and data information is collected for each 5 s. While doing testing in a case, the other cases are suspended (Fig. 7).

From these results above, when two virtual machines were put in the same compute node and same internal network, network performance was better than the other experiment cases. The measure results prove that virtual machines in the same compute nodes and same internal subnet get 4.2 % better results in network throughput when compared to the other cases.

On the other hand, packet delay time and packet loss between virtual machines in the same compute node but different internal subnets, or different compute nodes and different networks have better results. Especially, in case virtual machines are set up in the same compute node but different internal subnets has packet loss only 40 % when compared to other testing case with same compute node and same internal network (Fig. 8).

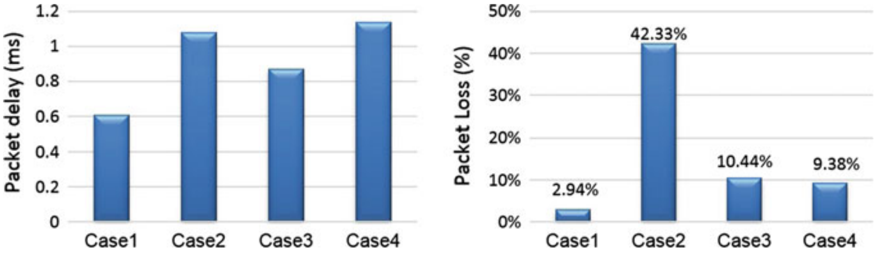


Fig. 8 Average packet delay time and packet loss in UDP protocol

4 Conclusion and Future Works

Cloud computing and open source cloud OpenStack today become a trend in technology and is one of the first choice of companies in the world. In this paper, we did a analysis and research on OpenStack network performance based on the parameters throughput, packet delay time, and packet loss. From the results, OpenStack Neutron ensure a high-performance network for virtual machines in cloud system. The results also prove that the different locations of virtual machine and internal network effect to network performance. For example, the virtual machines in the same compute node and same Internet subnet give the best network performance than other research cases.

However, in the paper, we only focus on UDP protocol in a small cloud system, this work needs to be research in a larger cloud system. In the future, our work will focus more on the other protocols (both UDP and TCP) and in a bigger cloud system with more test cases.

References

1. Borko Furht, Armando Escalante (2010), Handbook of Cloud Computing, Springer.
2. P. Mell and T. Grance. (2011). The NIST Definition of Cloud Computing.
3. Sabahi, F (2011). Cloud computing security threats and responses, 2011 IEEE 3rd International Conference on Communication Software and Networks (ICCSN).
4. Canonical. Ubuntu Cloud: Technologies for future-thinking companies. 2012. http://www.canonical.com/about-canonical/resources/white-papers/ubuntu-cloud-technologies-future-thinking-companies.
5. Cloud Security Alliance. About Cloud Security Alliance. https://cloudsecurityalliance.org/research/security-guidance.
6. OpenStack, http://www.openstack.org.
7. Shaoka Zhao et al (2013). Deployment and Performance Evaluation of Virtual Network based on OpenStack, in proceeding of CCIS 2013, pp. 18–22.
8. Qifeng Xu and Jie Yuan (2014), A Study on Service Performance Evaluation of Openstack, International Conference on Broadband and Wireless Computing, Communication and Applications (BWCCA), pp. 590–593.
9. F. Callegati, W. Cerroni, C. Contoli and G. Santandrea (2014), Performance of Network Virtualization in cloud computing infrastructures: The OpenStack case, 2014 IEEE 3rd International Conference on Cloud Networking (CloudNet), pp. 132–137.
10. Vo Nhan Van, Le Minh Chi, Nguyen Quoc Long, Nguyen Gia Nhu, and Dac-Nhuong Le (2015), A Performance Analysis of OpenStack Open-source Solution for IaaS Cloud Computing, in proceeding of International Conference on Computer and Communication Technologies (IC3T 2015), Advances in Intelligent Systems and Computing Vol.380, pp. 141–150 Springer.

A Shoulder-Surfing Resistant Graphical Password Authentication Scheme

M. Kameswara Rao, Ch. Vidya Pravallika, G. Priyanka and Mani Kumar

Abstract Many Internet-based applications authenticate the users before they are allowed to access the services provided by them. The traditional text-based password system is vulnerable to brute force attack, peeping attack, and reverse engineering attacks. The pitfalls of text-based passwords are addressed by employing graphical passwords. To create a password the graphical password systems make use of custom images, icons, or faces. Using images will decrease the tendency to choose insecure passwords. In this paper, we present a shoulder-surfing resistant pair based graphical password scheme to authenticate a user. Further enhancement of the proposed scheme is briefly discussed. Security analysis of the method is also evaluated.

Keywords Shoulder-surfing · Authentication · Graphical password · Security

1 Introduction

Confirmation figures out if a client ought to be permitted to access the services provided by a web-based application. Traditional passwords are utilized broadly for authenticating a client which has issues related to security and ease of use. Limitations of traditional authentication system include forgetting the password, selecting a guessable password, password theft, etc. So a major need to have a strong authentication system to secure all our applications is expected. Researchers turned out with cutting-edge mechanism called graphical password where they attempted to enhance the security and stay away from the shortcoming of traditional password system. Graphical passwords have been proposed as an option for

M. Kameswara Rao (✉) · Ch. Vidya Pravallika · G. Priyanka · M. Kumar
Department of Electronics and Computer Engineering, K L University, Guntur,
Andhra Pradesh, India
e-mail: kamesh.manchiraju@gmail.com

Ch. Vidya Pravallika
e-mail: passionatepravallika4u@gmail.com

H.S. Saini et al. (eds.), *Innovations in Computer Science and Engineering*,
Advances in Intelligent Systems and Computing 413,
DOI 10.1007/978-981-10-0419-3_13

traditional authentication system based on the fact that people can remember pictures superior to anything. Psychological studies have demonstrated that individuals can recollect pictures superior to anything [1] and recommends that people preferred perceiving visual data over reviewing text-based strings.

2 Related Works

Graphical password authentication schemes are categorized as recognition-based graphical password schemes and recall-based graphical password schemes. In recognition-based strategies, a client is given an arrangement of pictures and the client gets validated by perceiving and distinguishing the pictures, he or she chooses at the enrollment stage. In recall-based procedures, a client is requested to recreate something that he or she made at the registration stage.

2.1 Recognition-Based Graphical Password Schemes

Dhamija et al. [2] propounded a graphical password scheme in which the client is given an arrangement of irregular pictures from these pictures the client chooses a succession of pictures and for validation the client is told to recognize the pre-chosen pictures. Sobrado and Birget [3] built up a shoulder-surfing resistant graphical password scheme that showcase a blend of pass-articles (pre-chosen by client) and numerous different items. The scheme instructs the client to snap inside the convex hull space surrounded by all the pass-objects for verification.

In Man et al. [4] user chooses several pass-objects and the algorithm provides each of the pass-object with diverse variants and distinctive code. For verification, the client is given a few scenes where every scene contains a few pass-objects along with some fake objects. The client needs to type in a string with the codes of the individual pass-object variations distinguished in the scene. Jansen et al. [5] thesis has led to create an authentication scheme for mobile devices. During registration phase, a client chooses a topic (e.g., ocean, house and so forth.) that contains thumbnail photographs and afterward indicates a grouping of pictures for the password. For validation, the client must pick the enlisted pictures in a sequence. Takada and Koike [6] likewise recommended a comparative graphical password scheme for mobile devices that has a few phases of check for verification. This technique permits users for the usage of their favorite image to get authenticated. The clients first register their most loved pictures (pass-pictures) with the server. At every round, the client might either indicate a pass-picture among a few bait pictures or choose nothing if there is any pass-picture present. Real User Corporation submitted Passface Algorithm [7] in which the client needs to indicate four pictures of human appearances from an information base that contain faces for their future password. In the verification stage, the system displays a grid of nine confronts, a mix of one face already picked by the client and eight distraction faces.

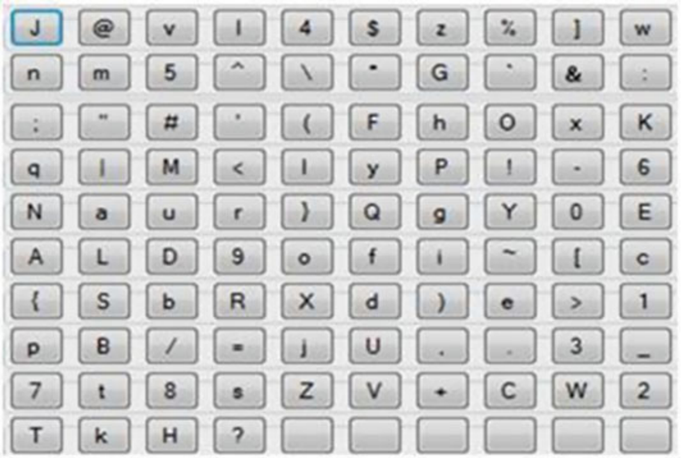
2.2 Recall-Based Graphical Password Schemes

Recreating a drawing and repeating a selection are the fundamental sorts of recall-based systems. Replicate a drawing strategy incorporates Draw-a-secret (DAS), Passdoodle system, Syukri strategy, and so forth. Jermyn et al. [8] proposed DAS method, which permits client to draw their remarkable password. A client is demanded to outline a photo on the 2D grid and the drawing's coordinates are saved in a succession. For authentication, the user has to re-draw the picture with the same coordinates in a specified order. Passdoodle method developed by Goldberg et al. [9] contains handwritten outlines generally drawn with a stylus on a touch screen. Syukri et al. [10] proposed a framework where verification is led by having client drawing their signature utilizing mouse. The system can modify the signature area and by enlarging or scale-down signatures and also by rotating if needed. The database stores this data and first takes the client information along with the signature's parameters for verification utilizing geometric normal means subsequent to performing the normalization once more.

In Repeat a Sequence of Actions group of authentication algorithms, a client is solicited to rehash arrangements from activities at first done by the client at the registration process. Strategies having a place with this classification incorporate Blonder system, Passpoint strategy, Passlogix technique, and so on. Blonder [11] developed a graphical password scheme in which a password is created by user tapping on different areas of a picture. During confirmation, the client must tap on the estimated territories of those areas. PassPoint Method proposed by Wiedenbeck et al. [12] that had developed Blonder's thought by evacuating the predefined limits and allowing distinctive arbitrary pictures to be utilized. Along these lines a client now can tap on any point of a picture for making a password and after that resilience about each chose pixel is ascertained. Keeping in mind the end goal to be authenticated, the client ought to click anyplace on the picture within the tolerance of the pixels that are chosen and also in a correct sequence. Passlogix method [13] has the bases of Blonder idea. For validation, the clients must tap on diverse things of the picture in the predefined grouping. Boundaries are characterized for everything undetectably so as to check if an item is chosen by mouse. Other related works can be found in [14].

3 Proposed Scheme

Our contribution is to propose a novel authentication scheme resistant to shoulder-surfing attack and discuss the key aspects of authentication for the proposed scheme. The proposed interface uses 94 characters, including A–Z, a–z, 0–9, and other printable characters as shown in Fig. 1. These characters are padded with spaces to form a 10 × 10 grid in which the characters are scattered randomly. User can select a password having at least 4 characters from the above 94 characters and



J	@	v	I	4	$	z	%	]	w
n	m	5	^	\	•	G	`	&	:
;	"	#	'	(	F	h	O	x	K
q	|	M	<	l	y	P	!	-	6
N	a	u	r	}	Q	g	Y	0	E
A	L	D	9	o	f	i	~	[	c
{	S	b	R	X	d	)	e	>	1
p	B	/	=	j	U	.	,	3	_
7	t	8	s	Z	V	+	C	W	2
T	k	H	?						

Fig. 1 The proposed graphical password interface

input his password by typing or by mouse clicks. User password is processed as pass-characters one pair at a time sliding to the right one character at a time wrapping around until the last pass-character forms the first element in the pair. Once the pass-characters are identified each pair is processed separately using predefined rules.

To illustrate the above process, let us follow an example where the user Alice's selects her password as "GUR@1". The pass-character pairs formed from the password are "GU", "UR", "R@", "@1" and "1G". The total number of characters in the password is equivalent to total number of pass-character pairs formed from the password.

Rules for processing the pass-characters

Rule 1: If both pass-characters appear on the same row of the grid, then the user must input any two characters in the row that lie between the two pass-characters in the pair included. For example, if the pass-character pair contains "A", "f" which appear on the same row of the grid then the user can input any of the two characters among "A", "L", "D", "9", "o", and "f".

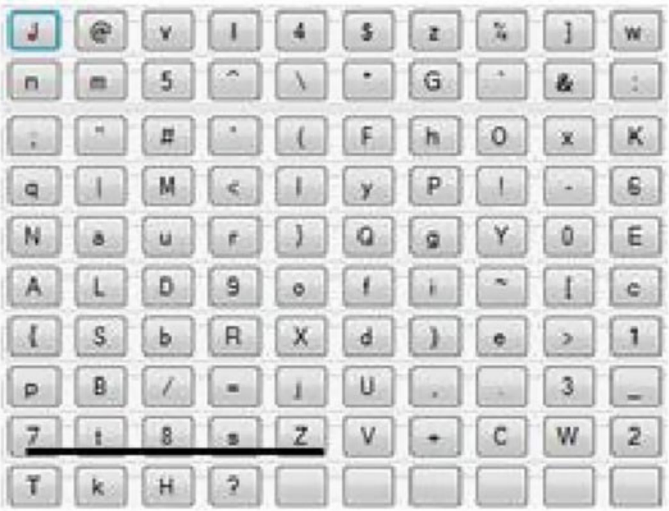
Rule 2: If the pass-characters appear on the same column of the grid, then input of the user must input any of the two characters in the column that lie between the two pass-characters in the current pair included. For example, if the pass-character pair contains "u", "8" which appear on the same column of the grid then the user can input any of the two characters among "u", "D", "b", "/", and "8".

Rule 3: If the pass-characters appear on different rows and columns, then input the characters on the corner points of the rectangle formed with the pass-characters as diagonal points. The rectangle corner on the same row as the first pass-character in the pair should be input first followed by the other rectangle corner. For example, if the pass-character pair contains "V", "3" which appear on different row and column then the user can input the characters on the corner points of the rectangle formed by "V", "3", i.e., "W" and "U", respectively.

Rule 4: If the two pass-characters are the same, they can be treated as appearing in the same row or column. In this case, the user can input any two characters surrounding the pass-character. For example, if the pass-character pair contains "J", "J" then the user can input any of the two characters among "n", "m", and "@".

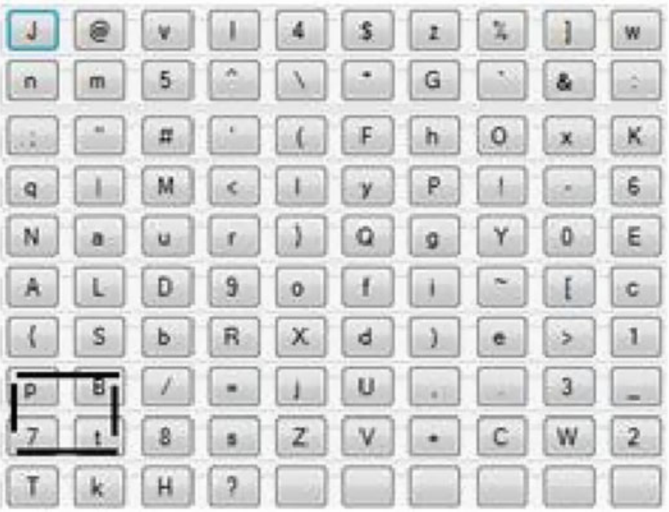
Let us consider the password selected by Alice as "Z7B4". The pass-character pairs formed are "Z7", "7B", "B4", and "4Z".

(1) Alice identifies the first pair of pass-characters "Z", "7" then as per Rule-1 she inputs any two characters that lie between "Z" and "7" (included) as identified in the Fig. 2.
(2) Alice finds the second pair of pass-characters "7", "B" then as per Rule-3 she inputs the characters "t" and "p" (the other corner characters) as shown in the Fig. 3.
(3) Alice finds the third pair of pass-characters "B", "4" then as per Rule-3 she inputs the characters "j" and "@" (the other corner characters) as shown in the Fig. 4
(4) Alice finds the fourth pair of pass-characters "4", "Z" and then as per Rule-2 she inputs any two characters between "4" and "Z" (included) as shown in the Fig. 5.



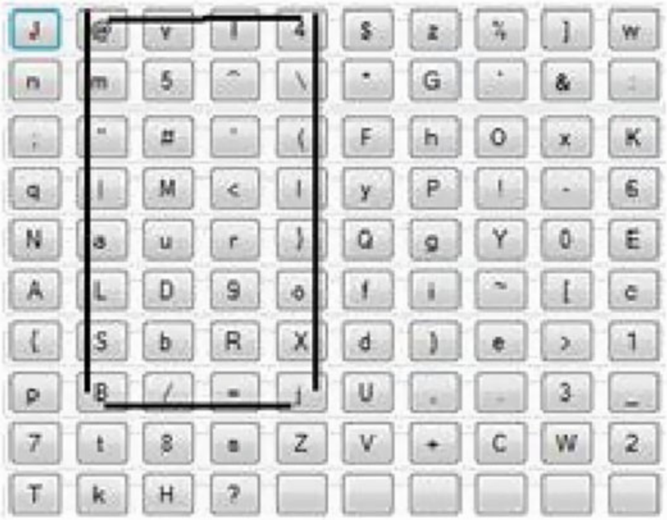
J	@	v	I	4	$	z	%	]	w
n	m	5	^	\	*	G	'	&	:
;	"	#	'	(	F	h	O	x	K
q	|	M	<	l	y	P	!	-	6
N	a	u	r	)	Q	g	Y	0	E
A	L	D	9	o	f	i	~	[	c
{	S	b	R	X	d	}	e	>	1
p	B	/	=	j	U	,	.	3	_
7	t	8	s	Z	V	+	C	W	2
T	k	H	?						

Fig. 2 First pair of pass-characters



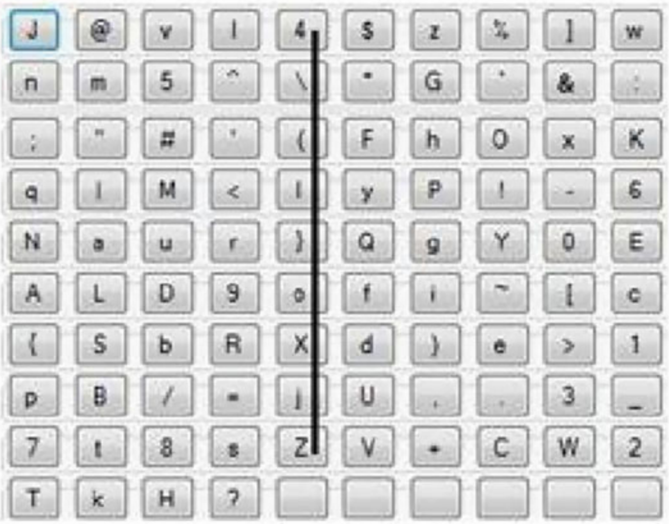
J	@	v	|	4	$	z	%	]	w
n	m	5	^	\	*	G	`	&	:
;	"	#	'	(	F	h	O	x	K
q	l	M	<	I	y	P	!	-	6
N	a	u	r	}	Q	g	Y	0	E
A	L	D	9	o	f	i	~	[	c
{	S	b	R	X	d	)	e	>	1
p	B	/	=	j	U	.	,	3	_
7	t	8	s	Z	V	+	C	W	2
T	k	H	?						

Fig. 3 Second pair of pass-characters



J	@	v	|	4	$	z	%	]	w
n	m	5	^	\	*	G	`	&	:
;	"	#	'	(	F	h	O	x	K
q	l	M	<	I	y	P	!	-	6
N	a	u	r	}	Q	g	Y	0	E
A	L	D	9	o	f	i	~	[	c
{	S	b	R	X	d	)	e	>	1
p	B	/	=	j	U	.	,	3	_
7	t	8	s	Z	V	+	C	W	2
T	k	H	?						

Fig. 4 Third pair of pass-characters



J	@	v	|	4	$	z	%	]	w
n	m	5	^	\	*	G	`	&	:
;	"	#	'	(	F	h	O	x	K
q	|	M	<	|	y	P	!	-	6
N	a	u	r	}	Q	g	Y	0	E
A	L	D	9	o	f	i	~	[	c
{	S	b	R	X	d	)	e	>	1
p	B	/	=	j	U	,	.	3	_
7	t	8	s	Z	V	+	C	W	2
T	k	H	?						

Fig. 5 Fourth pair of pass-characters

4 Security and Usability Study

A user study was conducted involving 18 B.Tech graduate students to study usability, security, and login times for the proposed scheme after a learning session on the proposed scheme. The average login time for the proposed scheme consisting of password length of 4, 5, and 6 password characters were 35.6, 44.2, and 49.3, respectively. The average login times increase as the password length increases in the proposed scheme. It was also found that 28 % of the participants in the study found the applying rules for the proposed scheme are time-taking when they choose a long password. Shoulder-surfing attack and dictionary attacks are restricted as the user is not directly typing the original password and the pass-characters are mapped with other input characters in the interface. In case of repeated attempts the security grid of the propose scheme will be changed. Since the password space in the proposed scheme is 94^n (n = number of password characters) which is more than the conventional text-based password space of 64^n. Hence the security of the proposed scheme is larger than the conventional text-based passwords.

5 Conclusions

A graphical password scheme is proposed to eliminate the shoulder-surfing attack and brute force attack. Shoulder-surfing attack is restricted as the user inputs other characters in place of the original password characters. When the number of login failures exceeds a certain threshold say 3 or 4, the interface changes the random order of characters thereby causing an adversary to start the attack from the beginning. This restricts random click attack in the proposed scheme. This work can be extended to cloud environment where the clients are authenticated using the proposed scheme before accessing the cloud services and they may also find existence in web login applications. The proposed scheme may find applications in mobiles which are facilitated with touch screens and also may be extended to ATM machines where the users be authenticated to log into their accounts. The proposed scheme can be extended using three color password characters (red, green, and blue) thereby increasing the password space by 3 × 94 = 282 characters.

References

1. E. Shephard, "Recognition memory for words, sentences, and pictures", *Journal of Verbal Learning and Verbal Behavior*, 6, pp. 156–163, 1967.
2. R. Dhamija and A. Perrig, "Deja vu: A user study using images for authentication", *Proceedings of 9th USENIX Security Symposium*, 2000.
3. X. Suo, et al., "Graphical passwords: A survey.", *Proceedings of 21st Annual Computer Security Applications Conference.*, pp. 463–472, 2005.
4. S. Man, et al, "A shoulder-surfing resistant graphical password scheme – WIW", *Proceedings of International Conference on Security and Management*, pp. 105–111, 2003.
5. W. Jansen, "Authenticating Mobile Device Users Through Image Selection", *Data Security*, 2004.
6. T. Takada and H. Koike, "Awase-E: Image-based Authentication for Mobile Phones using User s Favorite Images", *Human-Computer Interaction with Mobile Devices and Services*, 2795, Springer-Verlag GmbH, pp. 347–351, 2003.
7. Passfaces Corporation, "The science behind Passfaces", White paper, Available at http://www.passfaces.com/enterprise/resources/whitepapers.htm, July 2009.
8. A. Jermyn, et al., "The design and analysis of graphical passwords", *Proceedings of the 8th USENIX Security Symposium*, August, Washington, D.C., USA, 1999.
9. J. Goldberg, "Doodling Our Way to Better Authentication", *Proceedings of Human Factors in Computing Systems (CHI),* Minneapolis, Minnesota, USA.
10. A. F. Syukri, et al., "A User Identification System Using Signature Written With Mouse", *Third Australasian Conference on Information Security and Privacy (ACISP)*, Springer-Verlag Lecture Notes in Computer Science, pp. 403–441, 1998.
11. S. Chiasson, et al., "Graphical Password Authentication Using Cued Click Points", *ESORICS*, 24-27 September, Dresden,Germany, pp. 59–374, 2007.
12. S. Wiedenbeck, et al., "PassPoints: Design and longitudinal evaluation of a graphical pass-word system", *International Journal of Human-Computer Studies*, 63, pp. 102–127, 2006.
13. M. Boroditsky. Passlogix password schemes. http://www.passlogix.com.
14. R. Biddle, et al., "Graphical Passwords: Learning from the First Twelve Years", *ACM Computing Survey*, issue 44(4), 2011.

A Proposal for Searching Desktop Data

Mamta Kayest and S.K. Jain

Abstract Managing personal desktop data has become a necessity of the present day society as data on one's PC is increasing day by day. This data is huge as well as heterogeneous in nature. Users often need to locate the required data on desktop system. Therefore, how efficiently to find the required data items has become an emerging research issue. Various desktop search engines and tools are developed to provide search over the desktop data. In this paper, we propose a solution for managing heterogeneous desktop data.

Keywords Data · Desktop search engines · Metadata · Partial content retrieval

1 Introduction

The capacity of hard disk drives has increased tremendously; as a result, user stores a large number of files on his/her personal computer. So, certainly sometimes users face lot of difficulties in getting desired documents even though they know that they are saved somewhere on the disk. Nowadays, searching for documents can be faster on the World Wide Web than on our personal computer. Due to the availability of a variety of web search engines and ranking algorithms like the PageRank algorithm introduced by Google [1], web search has become more efficient than PC search. Therefore, there is a need of providing efficient search over desktop data to access required information. The main motivation of this work is to search files on desktop system efficiently for retrieving required data easily. Retrieval of partial information from files is also a necessity of users. In this paper, we have proposed diverse ways

M. Kayest (✉) · S.K. Jain
Computer Engineering Department, National Institute of Technology, Kurukshetra, India
e-mail: mamtakayest.nitkkr@gmail.com

S.K. Jain
e-mail: skj_nith@yahoo.com

H.S. Saini et al. (eds.), *Innovations in Computer Science and Engineering*,
Advances in Intelligent Systems and Computing 413,
DOI 10.1007/978-981-10-0419-3_14

of searching heterogeneous desktop data. The proposed system also retrieves partial contents from a semi-structured data files, e.g., XML files. Rest of the paper is organized as follows: Sect. 2 describes the related work; Sect. 3 discusses the proposed system for desktop data search; and finally, in Sect. 4 concludes the paper.

2 Related Work

Personal data refers to digital data accessed by a person during his/her lifetime and is owned by oneself. Personal data consists of heterogeneous data mix of word documents, pictures, XML data, audio file, video files, emails, and so on. This large amount of personal data may be spread on various devices like desktop system, laptop system, homepage server, e-mail server, official website, digital cameras, mobile phone, etc. For retrieving relevant information, effective management of personal data is required [2]. Desktop data is also personal data on one's desktop, but it is centralized in nature. Various desktop search engines (DSEs) have been developed for managing desktop data including Windows search [3], Google desktop search [1], Yahoo! [4], Corpernic Desktop Search [5], and many more, some of them are compared on various parameters in [6]. DSEs are based on the file systems of underlying operating systems and lack in capability of retrieving partial contents from files [7]. For searching through DSE approach, users first input search query to the search engine and then search engine transfer the query to the indexed database to get required result [8]. DSE employ one or more crawler programs on desktop files to crawl and extract information that are used by indexer to create an indexed database. Problem with DSEs are that they do not provide partial retrieval of information [7], no support for complex queries, no support for semantic integration, and take significant initial indexing time. Modeling and querying over heterogeneous desktop data is another important research issue. In iMeMex [9, 10] data model, a graph data model has been proposed for modeling personal data. A new Xpath-like query language named iQL is proposed to query over the uniform view, which is complex to understand by a novice as users are expected to have knowledge of the underlying structure of the personal data. Similarly, various methods are proposed to query over XML data [11–13].

3 A Proposal for Searching Desktop Data

This section discusses a solution for managing desktop data that includes various aspects of searching including metadata, relationships, and contents of XML files. The proposed work searches file system based on metadata and contents of semi-structured file. Figure 1 depicts a context diagram of the proposed system. Users input queries to the desktop search system, which in turn interacts with the file system for retrieving

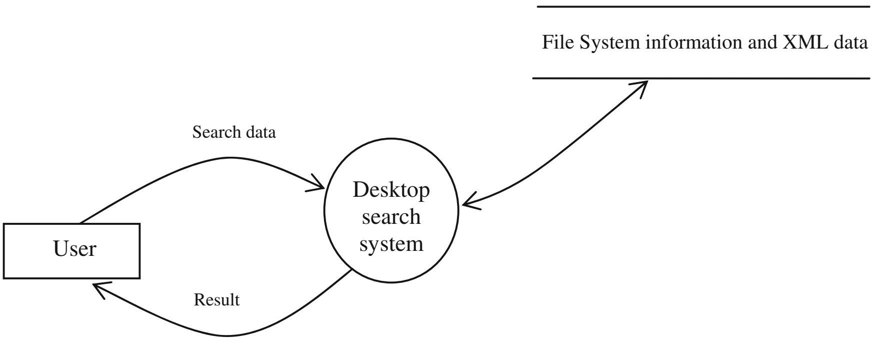


Fig. 1 Context diagram of the proposed desktop search system

necessary information. The system returns results to the user after processing queries. Figure 2 depicts a detailed DFD of the proposed desktop search system. The system is divided into two main modules; the first module makes search over files/folders based on their metadata and the second module process queries on XML documents. The proposed system offers options for making searches based on the metadata of files and folders, relationships, and contents of XML file. These options are summarized as follows:

- A file is searched based on metadata name, size, extension, and last modified date.
- A folder is searched based on metadata name, size, and last modified date.
- Relationship *hasfile* makes search on files.
- Relationship *hasfolder* makes search on folders.
- Retrieval of full contents of XML file.
- Retrieval of partial contents of XML files based on tags and field names.

For query over metadata of files/folders, first user enters the path of the file/folder and the relationship either *hasfile* or *hasfolder* to make search on files or folders. For example, a user searches all the files from drive *"d"* that were last modified on January 10, 2015. After giving path and relationship a hash table is created in memory containing various entries of file/folder's metadata and user gets result for files/folders based on the metadata as given in the query. This method of searching supports update guarantee as hash table is created in memory after entering the query. It also reduces the time taken for initial indexing of data by desktop search engines. Algorithm 1 makes search over files/folders based on their metadata and Algorithm 2 searches contents from XML files.

Algorithm 1 (Metadata-based search)

```
Step 1: enter the path and relationship
Step 2: map corresponding metadata entries in hash table: name, extension, size,
        last modified date for files and name, size, and last modified date for folders.
Step 3: if relationship is "hasfile"
```

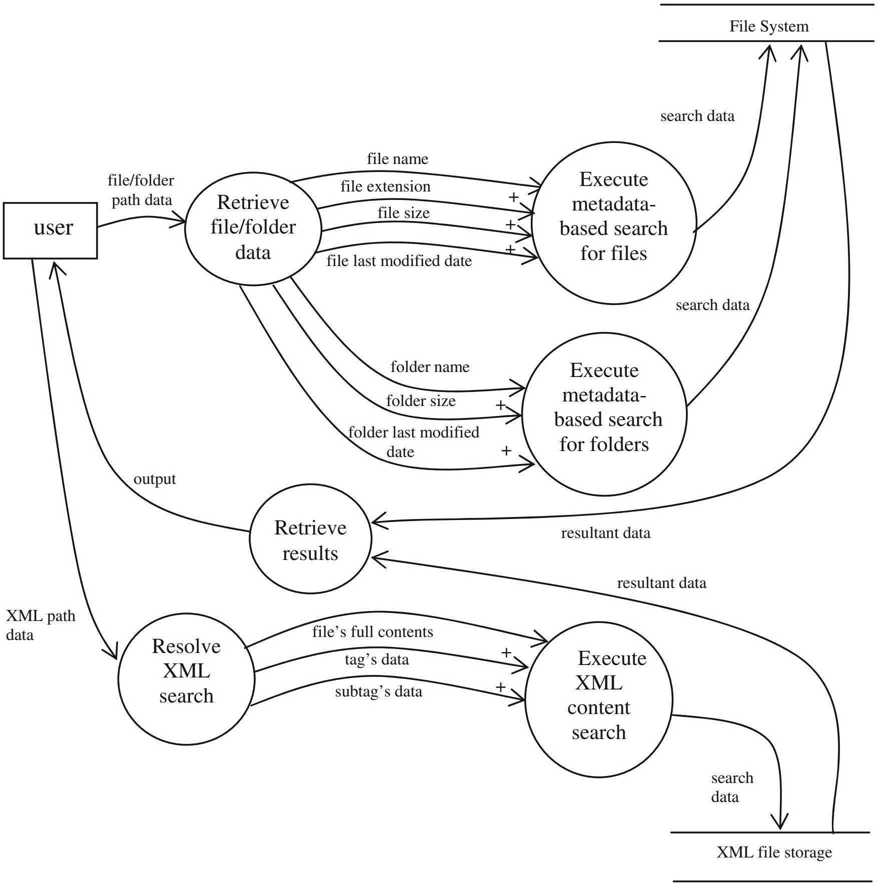


Fig. 2 Detailed DFD of the proposed work

```
         then
         read choice in ch for metadata from 1 to 4
         1. name 2. size 3. extension 4. last modified date
         else if relationship is "hasfolder"
         then
         read choice in ch for metadata from 1 to 3
         1. name 2. size 3. last modified date
         end if
Step 4:  if (ch == 1)
         then
         search hash map entry for file/folder name and print result
         else if (ch == 2)
```

```
          then
          search hash map entry for file/folder size and print result
          else if (ch == 3)
          then
          search hash map entry for file's extension/folder's last modified date and
          print result
          else if (ch == 4)
          then
          search hash map entry for file's last modified date and print result
          end if
```

Algorithm 2 (Content-based search on XML files)

```
Step 1:   enter path of file
Step 2:   read choice in ch for file's contents

          (1) full contents (2) tag's data (3) subtag/subfield's data

Step 3:   query parsed
Step 4:   if (ch == 1)

          then
          get and print full contents of XML file
          else if (ch == 2)
          then
          get and print all data of tag name
          else
          then
          get and print data of subtag
          end if
```

Some sample queries that the proposed system processes are

1. Search files from drive *d* where the file size is 500 MB.
2. Search file named *nisha* from *e* drive.
3. Search all folders from drive *g* which are modified on January 10, 2015
4. Search for folder named *nishafol* from *f* drive.
5. Search files from drive *d* which are modified on January 11, 2015.
6. Search all .xls files from *d* drive.
7. Display employee names from file EmpData.xml located in drive *d*.
8. Display employee's postal addresses from file EmpData.xml located in drive *d*.
9. Display all information related to employees from Empdata.xml, which is located in *f* drive.
10. Display contents of file *nisha.xml* from *g* drive.
11. Display employee's last names from file Empdata.xml from *d* drive.

4 Conclusion

Management of user's desktop system is a need of current society as desktop data is huge in amount and change frequently. Various desktop search systems such as Google, Corpenic, etc., are developed for management of personal desktop data. But these search engines require extra indexing time prior starting their work and also do not support partial retrieval of contents from files. In this paper, we propose design of a desktop data search system to which allows search over desktop data using metadata as well as partial and full content retrieval from files (XML files). The implementation of the proposed system is in its advanced stage, extending functionality of the proposed system in our future plan.

References

1. "Google Desktop" A desktop search engine from Google available at http://desktop.google.com, http://googledesktop.blogspot.in/ last visited on August 25, 2014.
2. Dittrich J. P., Blunschi L., Farber M. O. R. Girardm, Karakashian S. K., Antonio M., Salles V., "From Personal Desktops to Personal Dataspaces: A Report on Building the iMeMex Personal Dataspace Management System", proceedings of BTW 2007, 2007, pp. 292–308.
3. "Windows Desktop search" A desktop search engine from Microsoft available at http://www.microsoft.com/windows/products/winfamily/desktopsearch/default.mspx, last visited on December 29, 2014.
4. "Yahoo! Desktop Search" A desktop search engine from Yahoo available at http://info.yahoo.com/privacy/in/yahoo/desktopsearch/, last visited on December 1, 2014.
5. "Corpenic Desktop search" A desktop search engine from Microsoft available at http://www.copernic.com/en/products/desktop-search/home/download.html,last visited on Jan 9, 2015.
6. Markscheffel B., Buttner D., Fishcher D., "Desktop Search Engines- A State of the Art Comparison", proceedings of 6th International conference on Internet Technology and secured Transactions, 11–14 December 2011, pp. 707–711.
7. Pradhan S., "Towards a Novel Desktop Search Technique", proceedings of 18th International Conference on Database and Expert Systems Applications, DEXA 2007, held on September 3–7, 2007 at Regensburg, Germany, LNCS 4653, pp. 192–201.
8. Cole B., "Search engines tackle the desktop", IEEE, 2005.
9. Dittrich J. P., "iMeMex: A Platform for Personal Dataspace Management", proceedings on 2nd NSF sponsored Workshop on Personal Information Management, ACM SIGIR 2006.
10. Dittrich J. P., Salles M. A. V., "idm: A unified and versatile data model for personal dataspace management", proceeding of 32nd International Conference on Very Large Data Bases, VLDB 2006, held on September 12–15, 2006 at Seoul, Korea, pp 367–378.
11. Florescu D., Kossman D., Manolescu I., "Integrating keyword search into XML query processing" , proceedings of International World Wide Web Conference, pp. 119–135 (2000).
12. Pradhan S., "An algebraic query model for effective and efficient retrieval of XML fragments", VLDB, pp. 295–306 (2006).
13. Yunyao L., Cong Y., Hosagrahar J. V., "Schema-free XQuery", proceedings of 30th VLDB, pp. 72–83 (2004).

Low-Power Analog Bus for System-on-Chip Communication

Venkateswara Rao Jillella and Sudhakara Rao Parvataneni

Abstract At present, performance and efficiency of a system-on-chip (SoC) design depends significantly on the on-chip global communication across various modules on the chip. System-on-chip communication is generally implemented using a bus architecture that runs very long distances and covers significant area of the integrated circuit. The difficult challenges in design of a large SoC such as one containing many processor cores include routing complexity, power dissipation, hardware area, latency, and congestion of the communication system. This paper proposes an analog bus for digital data. In this scheme, it replaces 'n' wires of an 'n'-bit digital bus carrying data between cores with just one (or a few) wire(s) carrying analog signal(s) encoding '2^n' voltage levels. This analog bus uses digital-to-analog converter (DAC) drivers and analog-to-digital converter (ADC) receivers. This on-chip communication proposal can potentially save power and area. Diminution in the number of wire lines saves chip area and the reduction in total intrinsic wire capacitance consequently reduces the power consumption of the bus. The scheme should also reduce signal interference and cross-talk by eliminating the need for multiple line drivers and buffers. In spite of over-heads of the ADCs and DACs, this scheme provides significant power saving. Linear technology SPICE simulations show that the ratio of the power of the bus consumed by the proposed analog scheme to a typical digital scheme (without bus encoding or differential signalling) is given by $P_{analog}/P_{digital} = 1/(3n)$ where 'n' is the width of the bus.

Keywords System-on-Chip · Routing · Communication · Energy · Power

V.R. Jillella (✉)
Vignan Institute of Technology and Science, Hyderabad, India
e-mail: vraojillella@gmail.com

S.R. Parvataneni
Vignan Institute of Management and Technology for Women, Hyderabad, India
e-mail: sparvata@gmail.com

H.S. Saini et al. (eds.), *Innovations in Computer Science and Engineering*,
Advances in Intelligent Systems and Computing 413,
DOI 10.1007/978-981-10-0419-3_15

1 Introduction

In present ICs, power was a second order concern in chip design, succeeding the first order concerns of timing, area, testability, and cost. Nevertheless, for most system-on-chip (SoC) IC designs, low power dissipation is now one of the most momentous chip design purposes of any IC design. As power reduction is a product of overall improvement of the technology, it is not achieved through a single technological improvement. When the feature size is reduced down to the deep sub-micron region and power consumption is decomposed between the functional blocks and the communication paths between them, the power consumption has become a principal component. In relation with the multiple cores in the die, there is lack of literature on designing interconnect framework. Interconnect layout was done very late in the overall design because the conventional design flow was mostly logic based that emphasized on the design and optimization of logic. But at present as technology has enthused to gigahertz clock frequency and nanometer dimension, the design of interconnect plays a dominating role in determining performance, power, cost, and reliability [1–4]. There are very difficult defies in designing a large SoC, e.g., one comprising many processor internal cores, that includes routing complexity, power dissipation, hardware area, latency, and congestion of the communication network.

In present technology, recital and competence of SoC designs depend suggestively on the on-chip global communication across various modules on the chip. Generally the on-chip communication is implemented using a bus architecture that runs very long distances and covers a significant area of the integrated circuit. Rest of the paper is organized as follows. In Sect. 2, we have described the structure of analog bus. In Sect. 3, we evaluated the analog bus and the parallel bus. We concluded in Sect. 4.

2 Structure of Analog Bus

It has been appraised that DAC (digital-to-analog converter) and ADC (analog-to-digital converter) based inter-core communication structures considerably decrease the power consumption of various bit-line wide busses in multi-core computers and NOCs (networks-on-chip). The suggested scheme supplants an n-bit wide bus running between cores with a lone line, by coding the information (that needs to be carried out on n-bit bus) into 2^n levels of voltages on a single wire. Systems of this type offer the better of the two utmost prominent low power consumption inter-core communication systems—differential low voltage signalling and bus encoding, by encoding 'n' lines into one and keeping the low normal voltage signal swing. Diminution in number of signal wires and in total intrinsic wire capacitance consequently reduces chip area and power consumption. Additional advantages might comprise the removal of skew ambiguity due to

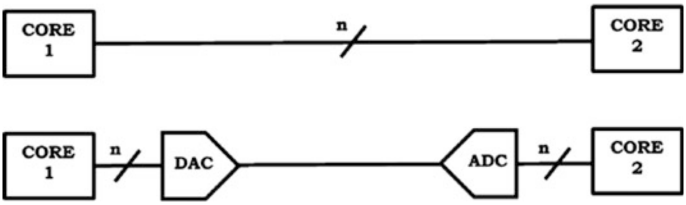


Fig. 1 Parallel bus and analog bus

elimination of various signal wires, timing verification simplicity, layout and blockage decrease due to abridged number of repeaters and vias. Such bus encoding can also be gainfully employed in test access mechanism for digital integrated circuits, as it could compress the total data to be communicated between the test head and core chip, thus reducing test time.

In our scheme 'n' wires of an n-bit digital bus carrying data between cores have been replaced with just one (or few) wire(s) carrying analog signal(s) encoding 2^n levels of voltage. For this, the analog bus utilizes DAC, (digital-to-analog converter) drivers and ADC, (analog-to-digital converter) receivers [5–7]. Figure 1 shows this transformation from 'n' wires (top) to a single wire (bottom). Such a system offers the finest of both the prominent low power inter-core communication systems—differential and low voltage signaling and bus encoding, in that it offers the eventual encoding 'n' lines to '1' and normal signal swing will be around $V_{\mathrm{DD}}/2$.

The power consumption of an analog bus would be,

$$P_{\mathrm{AnalogBus}} = V_{\mathrm{DD}} f \; V_{\mathrm{swing}} \, C \, \alpha$$

Generally the capacitance and supply voltage will remain same but there is reduction in the number of wires and voltage swing [5, 8–10].

3 Evaluation

In order to evaluate power reduction with the proposed analog bus scheme over a parallel bus, we first examine the power consumed in a case shown in Fig. 2, without any DAC/ADC elimination, using typical bus capacitances of large chips. Next, we examine the second case, shown in Fig. 3 where we replace the parallel

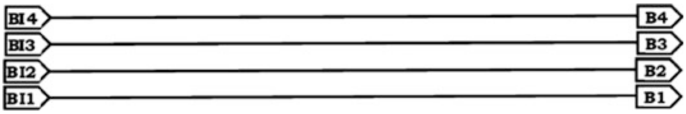


Fig. 2 4-bit parallel bus

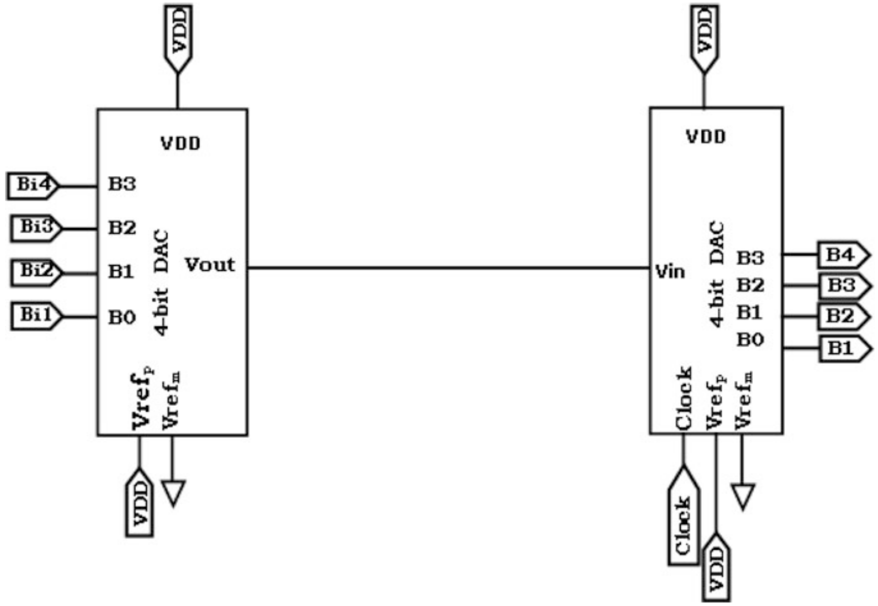


Fig. 3 Analog bus replacing 4-bit parallel bus of Fig. 2

Table 1 Experimental setup

Technology node	22 nm
Metal layer	4
Intermediate wire capacitance	2 pF/cm
Supply voltage	1 V
Simulation tool used	Linear technology SPICE
Spice models used	Ideal DAC and ADC
Activity factor	0.5
Frequency	500 MHz and 1 GHz
Input data pattern	Random
Wire length	1–5 mm

lines with a single line using ideal DAC and ADC from [11]. The simulations are done using simulation tool linear technology (LT) SPICE. LT SPICE is a high-performance SPICE simulation tool with enhancements and models for easing the simulation provided by linear technology (Table 1).

3.1 Experimental Setup for Power Analysis

Simulations have been done for two cases. First, a 4-line parallel bus has been replaced by a 1-wire analog bus, where both drive the same load circuit, a 2-bit adder. In the second case, an 8-line parallel bus has been replaced by a 1-wire analog bus, where both the setups drive a 4-bit adder.

3.2 Replacement of 4-Bit Parallel Bus by Analog Bus

For simulation, a 4-bit parallel bus has been replaced by a 1-line digital bus. This setup has been shown in Fig. 4. In this, the analysis has been made for bus lengths of 1–5 mm. Capacitance is calculated using the intermediate wire value given in the ITRS roadmap 2012 interconnect manual [12, 6, 13, 10, 14]. The digital input of the DAC is displayed in Figs. 5 and 6 displays the DAC output, which is transmitted to the ADC.

Comparison of power consumptions for frequencies of 500 MHz and 1 GHz with bus lengths of 1–5 mm (without addition of ADC/DAC power consumption) is given in Table 2. The average power consumption per mm for the analog bus is around 16.17 μW. The average power consumption per mm for each parallel line is 54.8 μW and for a 4-bit bus it is 219 μW for frequency of 500 MHz.

The average power consumption per mm for the analog bus is around 33 μW.

The average power consumption per mm for each parallel line is 115.8 μW and for a 4-bit bus it is 463 μW for frequency of 1 GHz.

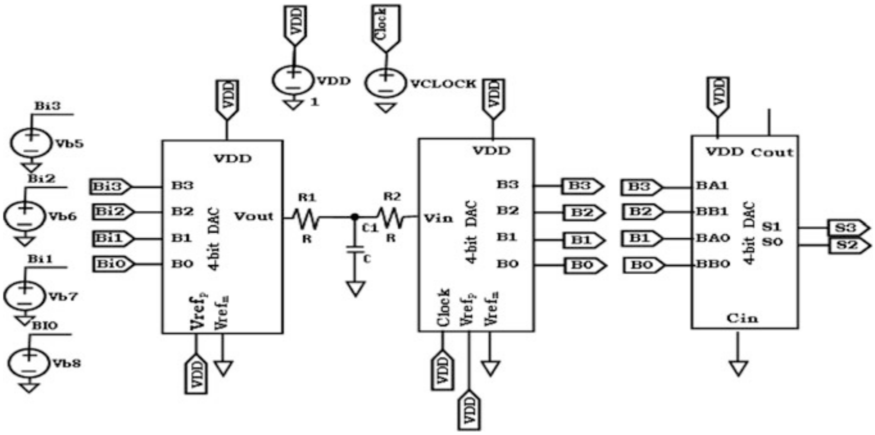


Fig. 4 Experimental setup for analog bus replacing a 4-bit parallel bus

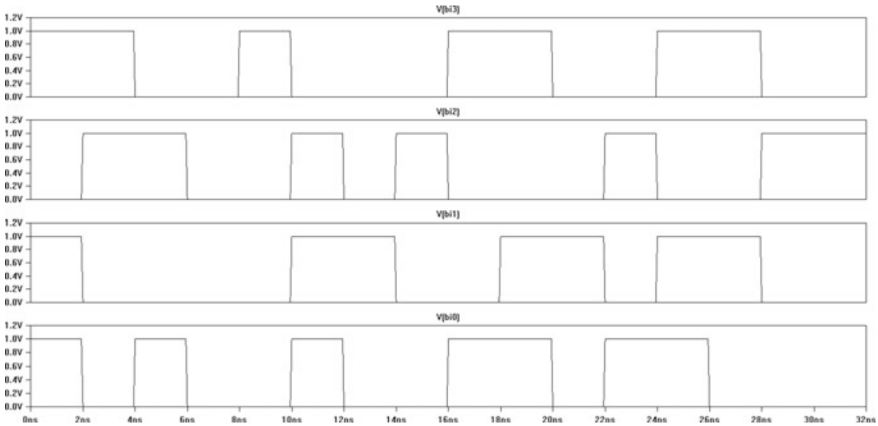


Fig. 5 4-Bit input patterns

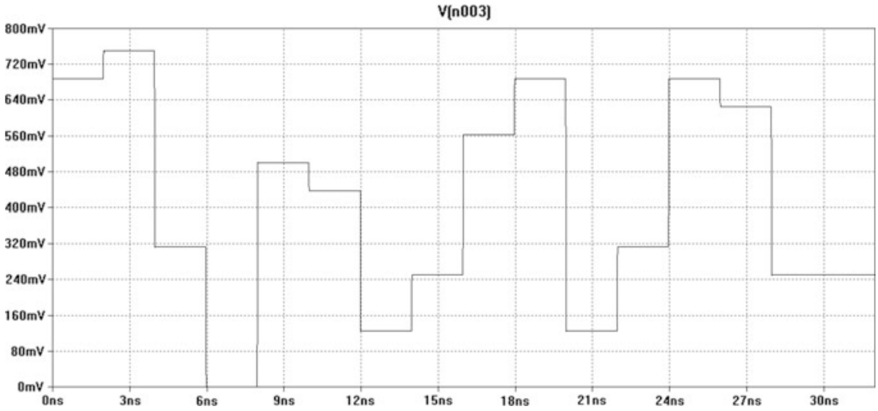


Fig. 6 4-Bit digital input converted to analog data

3.3 *Replacement of 8-Bit Parallel Bus by Analog Bus*

For simulation, an 8-line parallel bus has been replaced by a 1-line analog bus as shown in Fig. 7. The digital and analog signals are shown in Figs. 8 and 9 respectively. Here again the analysis has been done for bus lengths of 1–5 mm. Capacitance is calculated using the intermediate wire value given in the ITRS roadmap 2012 interconnect manual [12, 6].

Table 2 Comparison of power consumption of 4-bit parallel bus and analog bus for frequencies of 500 MHz and 1 GHz

Bus length (mm)	Parallel bus (μW) Freq. = 500 MHz	500 MHz analog bus (μW)	Parallel bus (μW) Freq. = 1 GHz	1 GHz analog bus (μW)
1	0219.22	018.3	0464.23	036.7
2	0438.95	033.73	0928.3	067.2
3	0658.13	046.87	01390	097.1
4	0875.34	059.28	01850	0126.5
5	01095	071.44	02310	0155.9

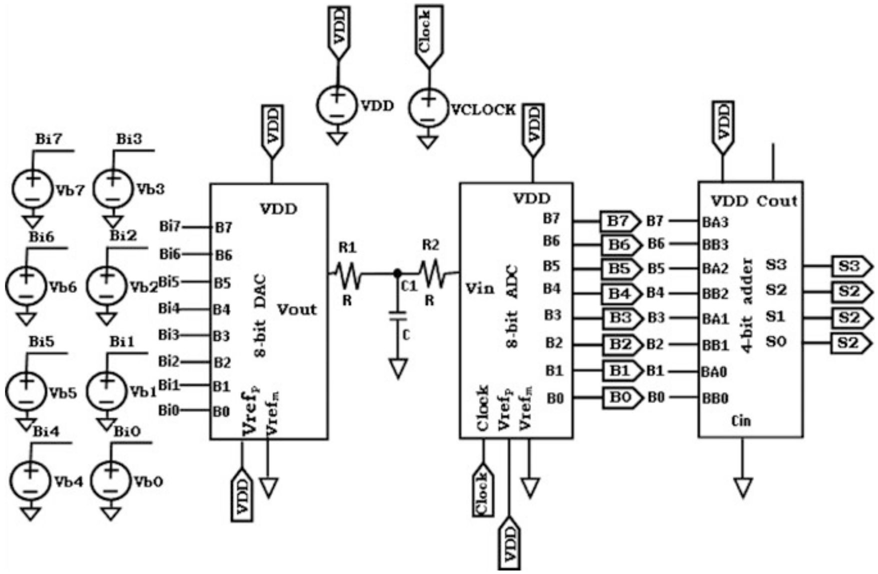


Fig. 7 An analog bus to replace an 8-bit parallel bus

Comparison of power consumption for a frequency of 500 MHz and 1 GHz with bus lengths of 1–5 mm (without addition of ADC/DAC power consumption) is given in Table 3.

The average power consumption per mm for the analog bus is around 18.3 μW. The average power consumption per mm for each parallel line is 58.65 μW and for an 8-bit bus it is 469.2 μW for 500 MHz. Table 4 gives the power saving of analog bus over 4-bit and 8-bit busses for frequency of 500 MHz and Fig. 10 shows the corresponding line graph.

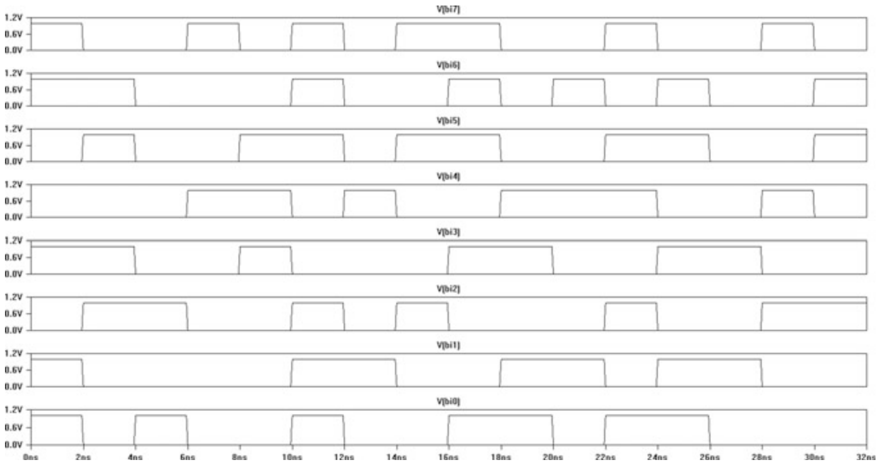


Fig. 8 8-bit input patterns

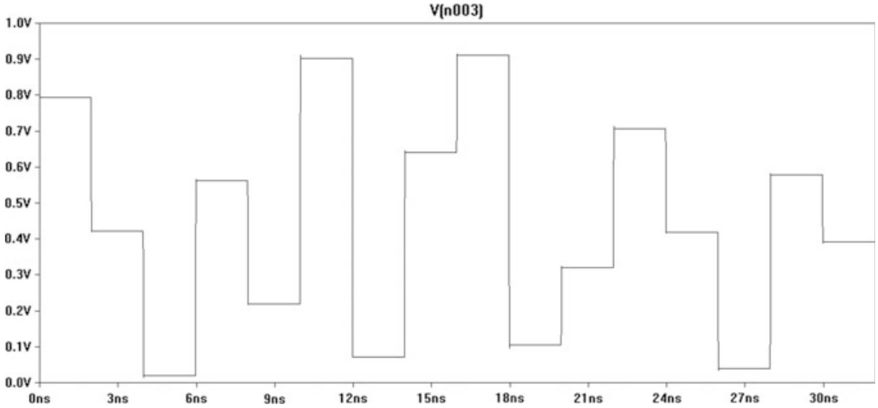


Fig. 9 8-bit digital input converted to analog data

Table 3 Comparison of power consumption of 8-bit parallel bus and analog bus for frequency = 500 MHz and 1 GHz

Bus length (mm)	Parallel bus (μW) Freq. = 500 MHz	500 MHz analog bus (μW)	Parallel bus (μW) Freq. = 1 GHz	1 GHz analog bus (μW)
1	0469.8	019.2	0995	038.50
2	0939	036.82	01990	073.35
3	01400	054.4	02980	0108.11
4	01880	071.84	03970	0142.59
5	02350	089.2	04950	0176.90

Table 4 Power saving of 4-bit and 8-bit busses for 500 MHz frequency

Bus length (mm)	4-bit bus power consumption			8-bit bus power consumption		
	Parallel bus (μW)	Analog bus (μW)	Power margin (μW)	Parallel bus (μW)	Analog bus (μW)	Power margin (μW)
1	219.22	18.3	200.92	469.8	19.2	450.6
2	438.95	33.73	405.22	939	36.82	902.18
3	658.13	46.87	611.26	1400	54.4	1345.6
4	875.34	59.28	816.06	1880	71.84	1808.16
5	1095	71.44	1023.56	2350	89.2	2260.8

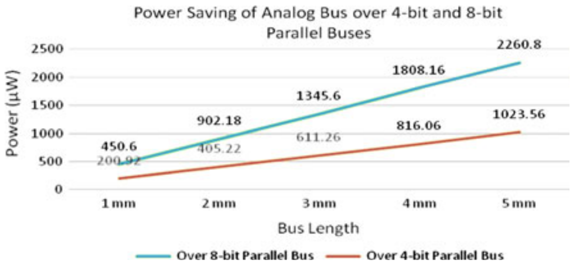


Fig. 10 Power saving of analog bus over 4-bit and 8-bit parallel busses for 500 MHz frequency

4 Conclusion

A distinctive concept of replacing parallel digital bus with an analog bus has been proposed here. A series of simulated experiments have been carried out to serve as proof-of-concept by evaluating power consumption of a single wire with DAC/ADC encoding in comparison to an n-bit parallel digital bus. The advantages of this scheme are reduced power consumption and reduced bus area, along with reduction of routing complexity, and congestion. LT SPICE simulation for an ideal case confirms that the ratio of bus power consumed by the proposed analog scheme to a typical parallel digital scheme (without bus encoding or differential signaling) is given by $P_{analog} = P_{digital} = 1/(3n)$, where n is the width of the bus.

References

1. J. Cong, "An Interconnect-centric Design Flow for Nanometer Technologies," Proc. IEEE,vol. 89, no. 4, 2001, pp. 505–528.
2. S. Pasricha and N. Dutt, On-Chip Communication Architectures: System on Chip Interconnect. Morgan Kaufmann, 2010.
3. G. E. Moore, "Lithography and the Future of Moore's Law," Proc. SPIE, vol. 2437, May 1995, pp. 2–17.

4. D. Ingerly, A. Agrawal, R. Ascazubi, A. Blattner, M. Buehler, V. Chikarmane, B. Choudhury, F. Cinnor, C. Ege, C. Ganpule, et al., “Low-k Interconnect Stack with Metal-Insulator-Metal Capacitors for 22 nm High Volume Manufacturing,” in Proc. IEEE International Interconnect Technology Conf., 2012, pp. 1–3.
5. A. D. Singh, “Four-valued Interface Circuits for NMOS VLSI,” International Journal of Electronics,vol. 63, no. 2, 1987, pp. 269–279.
6. Semiconductor Industry Association, “International Technology Roadmap for Semiconductors,” 2012.
7. “A Comparison of Network-on-chip and Buses,” White Paper, Arteris, SA, 2005.
8. L. Benini, G. De Micheli, E. Macii, M. Poncino and S. Quer, “Power Optimization of Core-Based Systems by Address Bus Encoding,” IEEE Trans. Very Large Scale Integration Systems, vol. 6, no. 4, 1998, pp. 554–562.
9. R. Ho, K. Mai and M. Horowitz, “Efficient on-chip global interconnects,” in IEEE Symp. onVLSI Circuits, 2003, pp. 271–274.
10. A. Kedia, “Design of a Serialized Link for On-chip Global Communication,” Master’s thesis, University of British Columbia, Canada, 2006.
11. B. Cordan, “An Efficient Bus Architecture for System-on-Chip Design“, in Proc. IEEE CustomIntegrated Circuits Conf., 1999, pp. 623–626.
12. http://www.itrs.net/Links/2012ITRS/Home2012.htm.
13. T. Bjerregaard and S. Mahadevan, “A Survey of Research and Practices of Network-on-Chip,”ACM Computing Surveys, vol. 38, Issue 1, 2006, Article No.1.
14. A. Kedia and R. Saleh, “Power Reduction of On-Chip Serial Links,” in IEEE International Symp. Circuits and Systems, 2007, pp. 865–868.

Text Clustering and Text Summarization on the Use of Side Information

Shilpa S. Raut and V.B. Maral

Abstract Clustering algorithm order information focuses on persuading social events concentrated around their similarity to abuse important data from data focuses. The end place of clustering these properties (text) has huge measure of information. It is difficult to measure relative data in light of the way in which the rate of the information is not clear. In such cases, it can be risky to partner side-data into the mining technique, since it can either build the nature of the representation for the mining system, then again add noise to the methodology. In various content mining applications, side-information is accessible nearby the content reports. Such text documents may be of a few sorts, for instance, record provenance information, the connections in the file, user access conduct from web logs, or other non-text based characteristics which are embedded into the content record. Such qualities may contain a massive measure of data for clustering purposes in the proposed system merge summarization methods. While executing the COATES estimation we used summarization system which is the union of duplicated clusters what's more, give last summary. COATES cluster algorithms we get the clusters on the establishment of substance what's more, auxiliary attributes. So in this project, an algorithm is designed, in order to give an effective clustering algorithm. Two algorithms are used in this project for clustering. In this paper COATES algorithm (this algorithm combines classical partitioning algorithms with probabilistic models) is used and the proposed system implements hierarchical algorithm which is compared with COATES algorithm and also implements the merging and summary generation algorithm which produces the summary or pure data for the user's convenience.

Keywords Clustering · Text mining · Auxiliary attribute · Clustering methods · Summarization

S.S. Raut (✉) · V.B. Maral
K. J. College of Engineering, Kondhwa 411043, Pune, India
e-mail: rautshilpa26@gmail.com

V.B. Maral
e-mail: vikasmaral@gmail.com

H.S. Saini et al. (eds.), *Innovations in Computer Science and Engineering*,
Advances in Intelligent Systems and Computing 413,
DOI 10.1007/978-981-10-0419-3_16

1 Introduction

The rapidly growing measures of text data in the setting of these broad online gatherings have driven an eagerness for making flexible and compelling mining algorithms. A colossal measure of work has been completed beforehand on the issue of clustering in text aggregations in the database and data recovery groups. Regardless of this, the work is essentially expected for the issue of impeccable text clustering, without diverse sorts of attributes. The clustering issue has of late been analyzed in connection of numeric information streams [1, 2]. In this paper, we will inspect this issue in the context of text likewise out and out information streams. Besides, the characteristic advancement [3] of stream data shows a couple of troubles to the clustering strategy. Most authentic applications routinely show experienced region which is not considered by most batch handling algorithms. The clustering issue presents different surprising troubles in a propelling data stream environment. For example, the steady improvement of clusters makes it essential to be prepared to quickly perceive new clusters in the data. While the clustering process needs to be executed reliably in online way, it is moreover basic to have the ability to give end clients the limit to separate the clusters in an online manner.

The worry of text clustering emerges in the setting of various application spaces, for instance, the Web and social networks. The rapidly extending measures of text information in the connection of these far-reaching online gradual additions have incited eagerness for making adaptable and successful mining algorithms. An enormous measure of work has been done as generally on the issue of clustering in content aggregations [2, 4–7] in the database and data recuperation groups. Regardless of this, the work has essentially got ready for the issue of unmodified text clustering, without diverse sorts of properties. In a few application spaces, a tremendous measure of side-information is also related close by the documents. This is because text documents regularly happen in the circumstance of a blend of employments in which there may be a far-reaching measure of distinctive sorts of database attributes then again meta-information which may be useful to the clustering process.

2 Related Work

Various applications, for instance, text crawling, news group filtering, and report association oblige continuous clustering, furthermore, division of text information records [8, 9]. The complete data stream clustering issue also has different applications to the issues of client division and ongoing pattern. We will demonstrate an online methodology for clustering tremendous text and supreme data streams with the usage of a measurable summarization technique. We present results representing the sufficiency of the method. Document clustering has not been by and large invited as a data recuperation instrument [6]. Dissents to its use fall into two crucial classes: the first is that clustering is greatly sensible for vast quantities (with running time consistently quadratic in the amount of records); also, second that clustering does not really upgrade recovery. Creators contradict that these issues emerge exactly when clustering is used as a piece of an attempt to upgrade routine pursuit procedures [4]. In any case, taking a look at clustering as a data access instrument in its own specific right blocks these complaints, and convinces to get a perfect model. We present a document searching method that adventures report clustering as its fundamental operation. We similarly present quick (straight time) clustering algorithms which offer assistance; this helps looking at the standard.

Spatial data mining is the disclosure of captivating associations; furthermore, attributes that may exist variably in spatial databases. In this paper, the creator mulled over whether clustering techniques have a measure to play in spatial information mining. To this end, the author produces another clustering technique called CLAHANS which is concentrated around randomized inquiry. Authors similarly create two spatial data mining algorithms that use CLAHANS [10]. The creator shows that with the assistance of CLAHANS, these two algorithms are extraordinarily intense and can brief disclosures that are difficult to find with current spatial data mining algorithms [11]. Furthermore, examinations coordinated to examine the execution of CLAHANS with that of existing clustering techniques show that CLAHANS is the most effective.

Finding significant samples in big datasets has attracted significant speculation; moreover, a champion between the most extensively measured issues here is the unmistakable verification of cluster, or populated districts, in a multidimensional dataset. Prior work does not address the issue of huge datasets and minimization of 1/0 costs. This paper shows a data clustering methodology named BIRCH (balanced iterative reducing furthermore clustering utilizing hierarchies), and demonstrates that it is especially fit for greatly vast databases. BIRCH incrementally and powerfully clusters approaching multidimensional metric data center to endeavor to pass on the best quality clustering with the available assets (i.e., accessible memory and time obligations) [12].

3 Implementation Detail

3.1 System Overview

The above figure shows that multiple documents give input to the preprocessing phase; all these documents contain the side-information. After that in preprocessing phase stemming and stop word removing is done for data to come into structured format. The latent semantic indexing is used for finding the cosine similarity between the documents, after that it gives document as input to the COATES algorithm which does the clustering for mining process (Fig. 1).

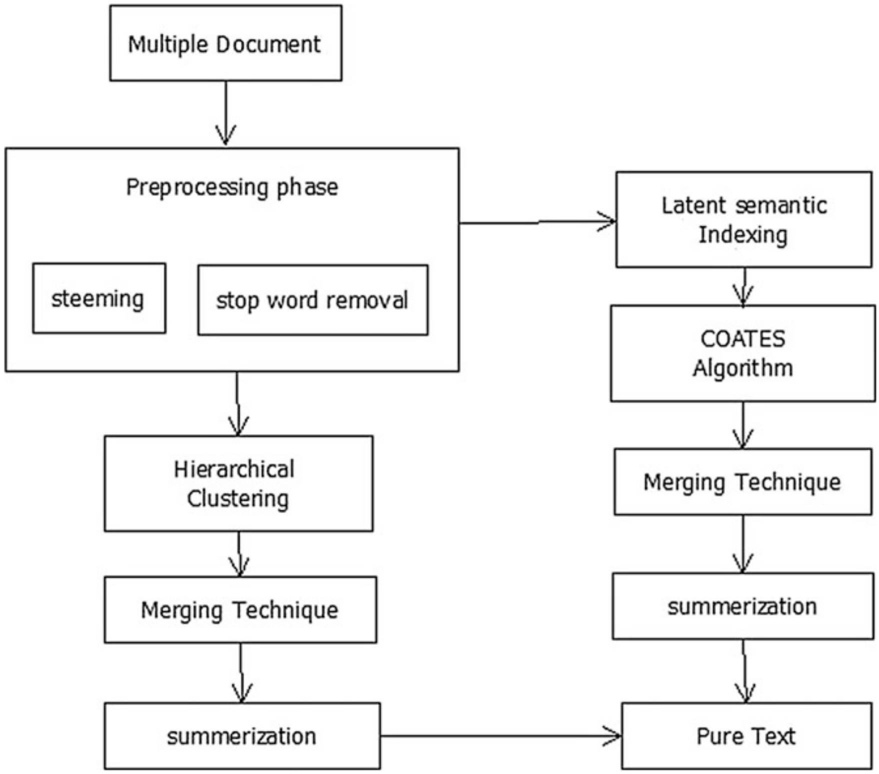


Fig. 1 System architecture

After that merging and summarization technique is applied for pure data. Then hierarchical clustering algorithm is implemented for clustering; this is done in the proposed system because it takes less time to do clustering. Therefore, the hierarchical algorithm is effective. Then implement merging and summarization technique on hierarchical clustering algorithm for giving pure data and compare both summaries and produce graph for accuracy, time, and memory.

3.2 Algorithms

Agglomerative Hierarchical Clustering Algorithm

```
Agglomerative Clustering ($D = \{x_i\}n$
$i = 1, k$):
$C = \{C_i = \{x_i\}| \, x_i \in D\}$ //Each point in separate cluster
$\Delta = \{\delta(x_i, x_j): x_i \, 2, x_j \in D\}$ //Compute distance matrix
while $|C| > k$ do
```

```
Find the closest pair of clusters $C_i$ 4, $C_j \in C$
5 $C_{ij} = C_i \cup C_j$ //Merge the clusters
6 $C = \{C - C_i - C_j\} \cup C_{ij}$ //Update the clustering
7 Update distance matrix $\Delta$ to reflect new clustering
```

Keyword summarization

Assume the key concepts K for a cluster C are known:

```
Step 1:  procedure SUMMARIZER($C$; $K$)
Step 2:  while $K$: size6 = 0 do
Step 3:  Rate all sentences in $C$ by key concepts $K$ (1)
Step 4:  Select sentence $s$ with highest score and add to $S$ (2)
Step 5:  Remove all concepts in $s$ from $K$ (3)
Step 6:  end while
Step 7:  return $S$
Step 8:  end procedure.
```

4 Results and Discussion

The following Table 1 shows k-means and COATES algorithm values for purity in percentage. The purity is obtained depending upon number of clusters (Figs. 2 and 3, Table 2).

Table 1 Comparison table for time

Clusters	COATES	Hierarchical
2	34	22
3	80	21
4	120	25
6	150	30

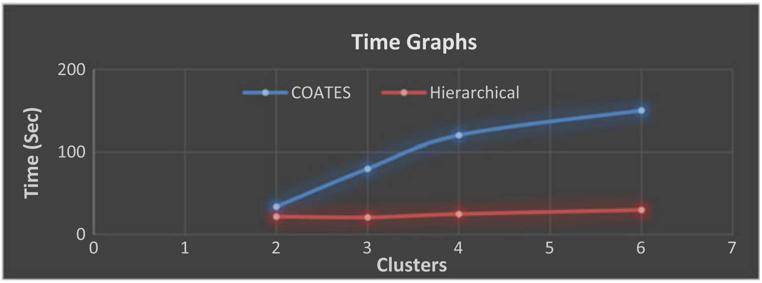


Fig. 2 Time comparison graph

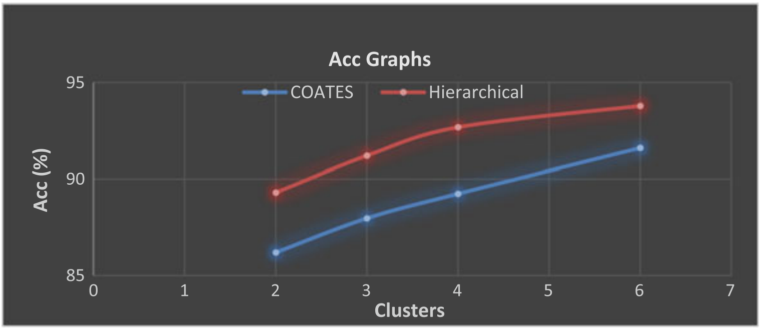


Fig. 3 Accuracy comparison graph

Table 2 Comparison table for accuracy

Clusters	COATES	Hierarchical
2	86.62	89.3
3	87.97	91.2
4	89.24	92.67
6	91.6	93.78

5 Conclusion

Mining text data with the utilization of side-data strategy is shown here. Various indications of text databases contain a considerable measure of side-information or meta-information, which may be used as a piece of appeal to advance the clustering system. To layout the clustering framework, we joined an iterative separating system with a probability estimation process which forms the vitality of different sorts of side-data. This general procedure is used inside request to framework both clustering and classification algorithms. The outcomes exhibit that the use of side-information can keep unplumbed the nature of text clustering and classification, while keeping up a strange condition of productivity. In the proposed framework we are arranging just to broaden this work utilizing summarization technique. Here in the wake of executing the COATES algorithm, we utilize merging summarization technique which is merges the reproduced cluster and gives last summary. In the wake of executing the COATES cluster algorithms we get the cluster on the premise of substance and auxiliary attributes. The merge technique takes basic substance and auxiliary attributes from this cluster furthermore, merge them and then apply summarization process for getting last summary.

References

1. D. Cutting, D. Karger, J. Pedersen, and J. Turkey, “Scatter/Gather: A cluster-based approach to browsing large document collections,” in Proc. ACM SIGIR Conf., New York, NY, USA, 1992, pp. 318–329.
2. R. Ng and J. Han, “Efficient and effective clustering methods for spatial data mining,” in Proc. VLDB Conf., San Francisco, CA, USA, 1994, pp. 144–155.
3. C. Aggarwal and P. S. Yu, “A framework for clustering massive text and categorical data streams,” in Proc. SIAM Conf. Data Mining, 2006, pp. 477–481.J.
4. S. Guha, R. Rastogi, and K. Shim, “ROCK: A robust clustering algorithm for categorical attributes,” Inf. Syst., vol. 25, no. 5, pp. 345–366, 2000.
5. D. Cutting, D. Karger, J. Pedersen, and J. Tukey, “Scatter/Gather: A cluster-based approach to browsing large document collections,” in Proc. ACM SIGIR Conf., New York, NY, USA, 1992, pp. 318–329.
6. W. Xu, X. Liu, and Y. Gong, “Document clustering based on nonnegative Matrix factorization,” in Proc. ACM SIGIR Conf., New York, NY, USA, 2003, pp. 267–273.
7. C. C. Aggarwal and P. S. Yu, “A framework for clustering massive text and categorical data streams,” in Proc. SIAM Conf. Data Mining, 2006, pp. 477–481.
8. I. Dillon, “Co-clustering documents and words using bipartite spectral Graph partitioning,” in Proc. ACM KDD Conf., New York, NY, USA, 2001, pp. 269–274.
9. Q. He, K. Chang, E.-P. Lim, and J. Zhang, “Bursty feature representation for clustering text streams,” in Proc. SDM Conf., 2007, pp. 491–496.
10. T. Liu, S. Liu, Z. Chen, and W.-Y. Ma, “An evaluation of feature selection for text is clustering,” in Proc. ICML Conf., Washington, DC, USA, 2003, pp. 488–495.
11. A. Banerjee and S. Basu, “Topic models over text streams: A study of batch and online unsupervised learning,” in Proc. SDM Conf., 2007, pp. 437–442.
12. S. Zhong, “Efficient streaming text clustering,” Neural Net w., vol. 18, no. 5-6, pp. 790–798, 2005.

Image Correspondence Using Affine SIFT Flow

P. Dedeepya and B. Sandhya

Abstract Image correspondence is one of the critical tasks across various applications of image processing and computer vision. The simple correspondence is being studied from many years for the purpose of image stitching and stereo correspondence. The images assumed for the simple correspondence have same pixel value even after applying the geometric transformations. In this work, we try to correspond images sharing similar content but vary due to change in acquisition like view angle, scale, and illumination. The use of features for flow computation was proposed in SIFT flow, which was used for correspondence across fields. In this method, dense SIFT descriptors are extracted and flow is estimated for matching the SIFT descriptors between two images. In this work, we applied SIFT flow algorithm on the affine transformations of images to be aligned.

Keywords Image correspondence · Geometric transformations · Scene alignment · SIFT flow

1 Introduction

The problem of image correspondence [1–3] deals with finding out the similar parts of two images with varying geometric, photometric, and temporal characteristics. Finding corresponding pixels or points between images becomes challenging when there is a change in view angle, scale, occlusion, etc. This has led to wide range research in image correspondence which can be broadly classified as feature based and pixel based.

P. Dedeepya (✉) · B. Sandhya
Department of Computer Science and Engineering,
M.V.S.R Engineering College, Hyderabad, India
e-mail: deepu.pothkanoori@gmail.com

B. Sandhya
e-mail: sandhyab16@gmail.com

H.S. Saini et al. (eds.), *Innovations in Computer Science and Engineering*,
Advances in Intelligent Systems and Computing 413,
DOI 10.1007/978-981-10-0419-3_17

Features and pixels have been used traditionally for sparse and dense correspondence respectively. However, features for the estimation of flow matrix have been first proposed in SIFT flow [4] to address the problem of scene alignment.

In this paper we have enhanced SIFT flow approach to address correspondence between images of the same scene but differing in terms of scale, rotation, and illumination effects. The paper is organized as follows: In Sect. 2 we present a survey of papers which have adopted or enhanced SIFT flow approach. In Sect. 3 we present the approach of Affine SIFT flow. In Sect. 4 we discuss the results obtained.

2 Related Work

SIFT flow proposed by Liu et al. [4] uses SIFT descriptor of 128-dimension for every pixel for the generation of the SIFT image. These SIFT descriptors are matched along the flow vectors keeping the flow field smooth. A dual-layer loopy belief propagation is used to optimize the matching objective. A coarse-to-fine approach is used in SIFT flow matching to improve the performance of the method.

In Table 1 we list the papers which have followed SIFT flow kind of approach for various applications.

We have adopted an approach similar to scale space SIFT flow, which is described in the following section.

3 Proposed Approach

SIFT flow inherits the merits of both the dense representation by obtaining pixel-to-pixel correspondences, and the sparse representation by matching transform-invariant feature of SIFT [12]. However, it is problematic in dealing with images with large change in geometric transformations and scale. To overcome this problem, we propose Affine SIFT flow approach. The pipeline of the approach is as shown in Fig. 1.

When matching two images with varying geometric transformations using the proposed method, Affine SIFT flow, we keep the second image unaltered and the first image is transformed into different affine transformations by varying the rotation [13, 14] and scale parameters. We use the scale field created by the scale space SIFT flow for finding the Affine SIFT flow. SIFT image is computed for the second image and transformed first image at different scales. These SIFT images are combined to form a single SIFT image for the first image.

Match the SIFT image computed at every geometric transformation of the first image at different scales with the SIFT image of the second image using SIFT flow and obtain the data term for every position at that geometric transformation. The flow at every position is computed between each pair of the second SIFT image and

Table 1 Survey of SIFT flow based approaches

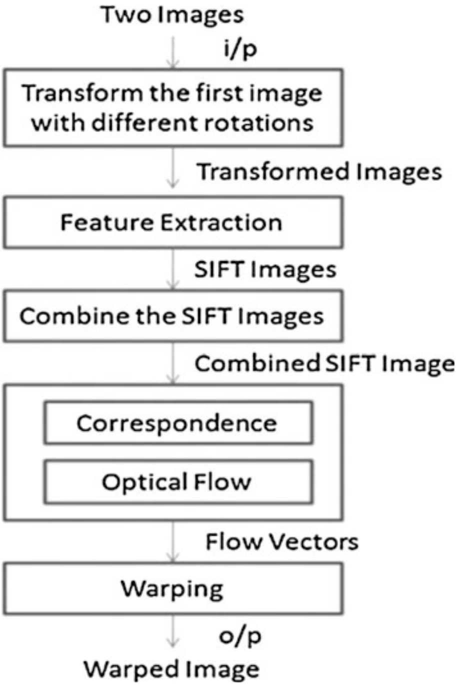
References	Title	Method	Applications
[4]	SIFT flow: dense correspondence across scenes and its applications	Dense SIFT descriptors are matched along the flow vectors using dual-layer loopy belief propagation to optimize the matching objective	Motion field prediction from single image, motion synthesis via object transfer, face recognition
[5]	Nonparametric scene parsing: label transfer via dense scene alignment	Warps the existing annotations and integrates multiple cues in a Markov random field framework to segment and recognize the query image	Object recognition and scene parsing
[6]	Non-rigid dense correspondence with applications for image enhancement	The nearest neighbor for each patch of the source image is found in the reference images, searching over a constrained range of translations, scales, rotations and bias values	Local color transfer, image deblurring and mask transfer
[7]	Understanding discrete facial expressions in video using an emotion avatar image	Adopted SIFT flow for aligning the face images, which is able to compensate for large rigid head motion and maintain facial feature motion detail	Facial expression recognition and analysis
[8]	Dense image correspondence under large appearance variations	Instead of matching pixels at pre-defined feature scales and rotations, they are treated as unknown variables that this method tries to solve for	Structure-from-motion, image retrieval, and object recognition
[9]	Dense correspondences across scenes and scales	Attempt to produce robust, dense descriptors by treating each pixel independently without considering the scales of other pixels in the image	Single-view depth estimation, semantic labels and segmentation, image labeling
[10]	On SIFTs and their scales	Each pixel is represented by a set of SIFT descriptors extracted at multiple scales and matched from one image to the next using set-to-set similarities	Object recognition

(continued)

Table 1 (continued)

References	Title	Method	Applications
[11]	Scale-space SIFT flow	Matching is done by computing SIFT feature at every scale of every pixel in the first image keeping the second image at its own scale. The best match is estimated with the selection of right scale factors through minimizing feature matching cost, keeping the flow field smooth and keeping the scale field smooth	Scene parsing, image/video retrieval, motion estimation and depth estimation

Fig. 1 Pipeline of the system



the transformed SIFT image of the first image to obtain flow vectors as in the SIFT flow. Combine the flow by minimizing the flow field of the images at every position keeping the flow field and the scale field smooth to obtain flow vectors for warping. Warp the flow image with the original image using flow vectors obtained from combining the flow and minimizing the flow to get the warped image.

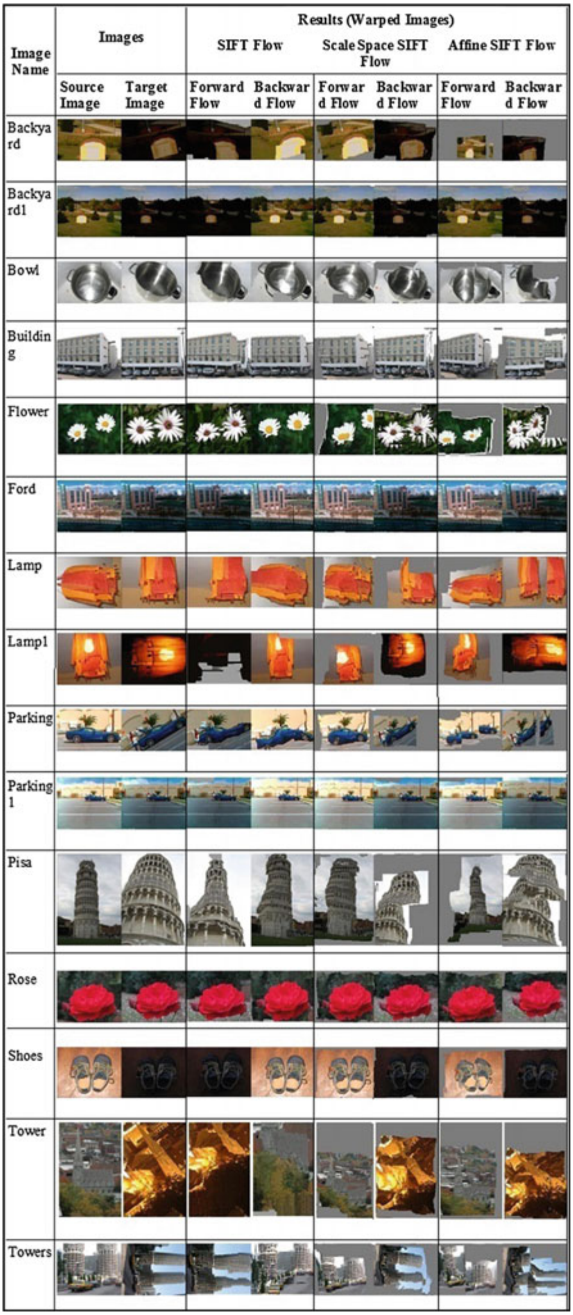


Fig. 2 Warped images using the SIFT flow, scale space SIFT flow and Affine SIFT flow

4 Results and Discussions

We have used SIFT flow library [15] in implementing our method, Affine SIFT flow. The input image is transformed into 11 transformed images. These transformed images are formed by rotating the source image with a difference of 30°, i.e., 30°, 60°, 90°…330°. SIFT images for the transformed images are scaled with scales 1, 2, 4, 6, and 8. Flow between the SIFT image of the target image and the SIFT images of the transformed images is computed using the dual-layer belief propagation for optimizing the flow vectors.

The images taken for testing this method are taken from the SLS Dataset [10], INRIA Holiday Dataset [16], GTILT Dataset [17] and Challenging Image Pairs [18]. Figure 2 shows the results, i.e., warped images for the source and target images selected from the datasets, for the three methods namely SIFT flow, scale space SIFT flow and Affine SIFT flow. The time taken for execution of these methods is mentioned in Table 2.

From Table 2, we get to know that the time of execution for our method Affine SIFT Flow is large and hence needs to be optimized.

Table 2 Time of execution for the methods SIFT flow, scale space SIFT flow and Affine SIFT flow

Image names	Image size	Time of execution in seconds					
		SIFT flow		Scale space SIFT flow		Affine SIFT flow	
		Forward flaw	Backward flaw	Forward flaw	Backward flaw	Forward flaw	Backward flaw
Backyard	250 × 150	10.04	10.26	102.63	103.69	303.70	295.18
Backyard1	224 × 168	10.34	10.91	102.27	101.66	303.47	303.78
Bowl	250 × 150	10.07	10.05	102.93	103.63	301.31	300.62
Building	192 × 144	7.47	7.51	76.41	76.08	217.10	217.32
Flower	200 × 152	8.19	8.66	79.58	79.48	237.28	238.12
Ford	192 × 144	7.42	7.46	74.84	75.03	213.96	215.45
Lamp	198 × 148	8.01	7.85	80.55	79.77	227.44	226.68
Lamp1	198 × 148	7.60	7.74	77.86	78.82	229.28	226.50
Parking	250 × 150	10.01	10.51	103.32	103.02	293.64	296.44
Parking1	192 × 144	7.35	7.53	75.20	75.66	212.11	21.93
Pisa	154 × 205	8.65	8.70	85.78	85.53	247.21	246.48
Rose	240 × 180	12.07	12.28	117.32	116.83	334.61	334.7
Shoes	208 × 156	8.76	8.66	88.27	87.38	250.30	249.59
Tower	166 × 250	11.27	11.38	115.15	125.17	330.12	331.98
Towers	208 × 156	8.77	9.54	87.82	87.87	248.64	250.87

5 Conclusion and Future Work

From the results, we conclude that the images with large differences in the viewpoints and scales give comparable results with the scale space SIFT flow and SIFT flow methods. With our method we explored a possibility of using SIFT flow to effectively deal with image pairs having large differences in viewpoint and scales.

This work can be extended in future by optimizing the computational complexity and memory consumption. We can test this method with other feature descriptors [19, 20] like SURF, GLOH, etc.

References

1. Barbara Zitova and Jan Flusser, Image Registration Methods: A Survey, Image and Vision Computing, 2003.
2. Abhijit, S. Ogale and Yiannis Aloimonos, Shape and the stereo correspondence problem, International Journal of Computer Vision, 2005.
3. Siddharth Saxena and Rajeev Kumar Singh, A Survey of Recent and Classical Image Registration Methods, International Journal of Signal Processing, Image Processing and Pattern Recognition, 2014.
4. Ce Liu, Jenny Yuen and Antonio Torralba, SIFT Flow: Dense Correspondence across Scenes and its Applications, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011.
5. Ce Liu, Jenny Yuen and Antonio Torralba, Nonparametric Scene Parsing via Label Transfer, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011.
6. Yoav Ha Cohen, Eli Shechtman, Dan B Goldman and Dani Lischinski, Non-Rigid Dense Correspondence with Applications for Image Enhancement, Association for Computing Machinery, 2011.
7. Songfan Yang and Bir Bhanu, Understanding Discrete Facial Expressions in Video Using an Emotion Avatar Image, IEEE Transactions on Systems, Man and Cybernetics, 2012.
8. Linlin Liu, Kok-Lim Low and Wen-Yan Lin, Dense Image Correspondence Under Large Appearance Variations, International Conference on Image Processing, 2013.
9. Moria Tau and Tal Hassner, Dense Correspondences Across Scenes and Scales, arXiv:1406.6323v1 [cs.CV], 2014.
10. Tal Hassner, Viki Mayzels and Lihi Zelnik-Manor, On SIFTs and their Scales, IEEE Transaction on Computer Vision and Pattern Recognition, 2014.
11. Weichao Qiu, Xinggang Wang, Xiang Bai, Yuille A and Zhuowen Tu, Scale Space SIFT Flow, IEEE Conference on Applications of Computer Vision, 2014.
12. D. G. Lowe, Object Recognition from Local Scale-Invariant Features, IEEE International Conference on Computer Vision (ICCV), 1999.
13. Dmytro Mishkin, Michal Perdoch and Jiri Matas, Two View Matching with View Synthesis Revisited, Center for Machine Perception, 2013.
14. Jean-Michel Morel and Guoshen Yu, ASIFT: A new Framework for Fully Affine Invariant Image Comparision, Image Processing On Line, 2011.
15. SIFT Flow Matlab code - http://people.csail.mit.edu/celiu/ECCV2008/.
16. D. Martin, C. Fowlkes, D. Tal, and J. Malik, A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics, Conference of Computer Vision, 2001.
17. M.R. Bales, D. Forsthoefel, D.S. Wills, and L.M. Wills, “The Georgia Tech Illumination Transition (GTILT) Dataset”, Mobile Vision Embedded Systems Lab (MoVES), 2012.

18. G. Yang, C.V. Stewart, M. Sofka and C.L. Tsai, Registration of Challenging Image Pairs: Initialization, Estimation, and Decision, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007.
19. Fabio Bellavia, Domenico Tegolo, Improving SIFT-based descriptors stability to rotations, Intenational Conference on Pattern Recognition, 2010.
20. E. Tola, V. Lepetit and P.Fua, DAISY: An Efficient Dense Descriptor applied to Wide-Baseline Stereo, IEEE Transaction on Computer Vision and Pattern Recognition, 2010.

Link Quality-Based Multi-hop Relay Protocol for WiMedia Medium Access Control

K.S. Umadevi and Arunkumar Thangavelu

Abstract WiMedia Alliance have been widely adopted in personal area networks for high data transfer with low energy consumption-based applications. The multi-hop support provided by WiMedia is limited to 3-hop distance and seeks significant interest from researchers. This paper is aimed at addressing the following issues by two-fold approach. In the first part, we propose a multi-hop routing algorithm for n-hop support to analyse the protocol and achieved higher throughput and minimum delay. Further, we tested the performance of the algorithm by implementing the protocol to quantify the outcomes. The results indicate that the proposed algorithm facilitates better utilisation of resources and helps to improve the network life time by providing high data rate.

Keywords Distributed network · WiMedia MAC · Multi-hop · Link quality · Throughput

1 Introduction

Wireless personal area network is a prominent technology in personal area network having tremendous growth in the last decade due to its communication opportunities and better quality of service. In fact, there is a high level of expectation to support multimedia applications, constructing home networks, military applications, medical applications but the performance is limited by its coverage region. Its operating frequency range starts from 2.4 GHz and extends to 10.6 GHz. When devices come into close proximity, they can communicate with their neighbours synchronised in the network.

K.S. Umadevi (✉) · A. Thangavelu
School of Computing Science and Engineering, VIT University,
Vellore, India
e-mail: umadevi.ks@gmail.com

A. Thangavelu
e-mail: arunkumar.thangavelu@gmail.com

H.S. Saini et al. (eds.), *Innovations in Computer Science and Engineering*,
Advances in Intelligent Systems and Computing 413,
DOI 10.1007/978-981-10-0419-3_18

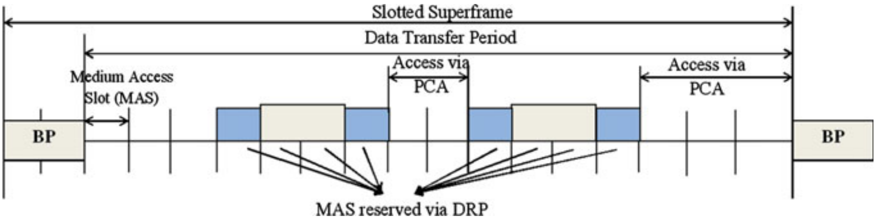


Fig. 1 Superframe structure

WiMedia technology uses multiband orthogonal division modulation (MB-OFDM) to support short-range high-bandwidth communication at low energy levels using ultra-wide band physical layer [1]. When compared to the other personal area network protocols, the success of WiMedia is the capacity of providing maximum data rate of 480 Mbps with low power consumption. It provides synchronized and distributed access, which is simple and scalable.

WiMedia uses Superframe structure for medium access (Fig. 1), which encompasses beacon period (BP) for synchronisation and data transfer period using two modes, namely reservation-based access using Distributed Reservation Protocol (DRP), contention-based access using Prioritised Contention Access (PCA). WiMedia MAC enables the node(s) to transmit a packet only when it is synchronized with the receiving node using beacon frames then identify the freely Medium Access Slot using the same set of frames either through beacon frames or control frames. Using beacon frames, the source node will be able to identify the link quality indicator (LQI) and the supportable data rate of the receiving node. If no free slots are available, it may try to send the data using PCA mode.

In Sect. 2, we have discussed about the need for relaying and proposed the procedure for identifying relay nodes. In Sect. 3, we propose a relay format for dealing with adding/updating the new relay request. In the rest of the paper, we have analysed the impact of relay node using the quality of service metrics i.e. delay and throughput.

2 Link Quality-Based Relay Selection

In wireless network, the strength of the network is purely based on transmission medium used by the signals. Signals may be deteriorated due to interference, network topology and power consumption. These factors may affect network throughput and lifetime. So to avoid problems that result due to the above mentioned factors, multi hopping is suggested [2–5] where the source and destination nodes use intermediate nodes, called cooperative relays to continue the data transfer. In turn, the use of cooperative nodes may reduce the packet drop ratio and increase the throughput [2]. In multi-hop wireless networks, cooperative relay nodes are used to utilise the idle slots [3]. The relay nodes are selected based on the acceptable link capacity between source to relay and relay to destination pairs.

Beacon frames are used for broadcasting information about device capacity and the details about registered and unregistered slots. These details include Link Quality Indicator IE (LQI IE), Distributed Reservation Protocol IE (DRP IE) and so on. The path between any two nodes can be finalised by identifying the free slots using DRP IE [6, 7]. Using Link Quality Indicator/Receiver Signal Strength Indicator, signal strength supported between source to relay and relay to destination are identified [6]. If the identified signal strength is less than lower threshold value (SIR threshold > 5.3 dB) then to continue communication, the sender needs to change the data rate [8]. With the help of DRPIE, free slots are identified between source to destination, source to relay and relay to destination pairs (Fig. 2).

Using DRP, the paths between source and destination (106.7 Mbps), source and relay (200 Mbps), relay and destination (320 Mbps) are finalised with the available data rates using either beacon frames or control frames [6, 9].

The signal strength between source and destination is less. So source node may prefer a relay for forwarding data packets to achieve a higher data rate. For WiMedia networks, to the best of our knowledge, researchers proposed only 3-hop distance relaying [9]. The same concept can be extended for multiple relay nodes (h), say if n nodes are preferred then h_{i} and $h_{\mathrm{i}+1}$ should agree on their link usage level where $1 \leq i \leq n$. So in this paper, we are interested in proposing a multi-hop relay protocol for WiMedia using link capacity to support n number of nodes in order to enhance the throughput and network lifetime. The objective of this work is to

- Identify relay node(s) with good link quality using data rate.
- Create a scalable network to extend network life time.

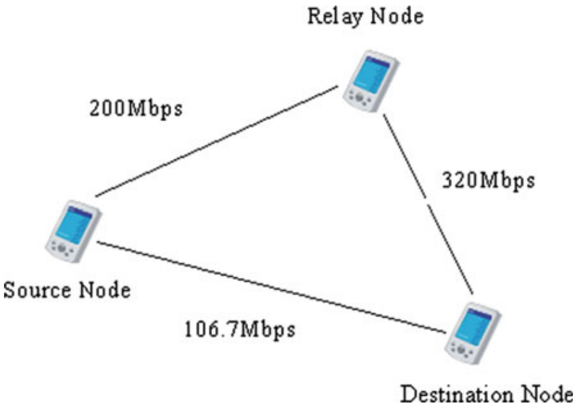


Fig. 2 Choosing relay node

3 Link Quality-Based Multi-hop Relay Protocol for WiMedia Mac

In wireless networking environment, the data rate of a node indicates many factors like delay, throughput, energy level and packet drop rate which in turn affects network lifetime. The best link quality optimises the power consumption, transmission rate, and throughput. If the link quality is lowered through interference or low residual power or not in coverage proximity then source node must identify the relay by using link quality. In Fig. 3, source node identifies the destination node not in coverage region and hence chooses a relay node. Reaching destination is possible via R_1, R_2 and R_3. Hence source node identifies R_1 as the relay node using link quality where relay node is selected based on maximum of link quality between {source node and R_1, source and R_2, source and R_3}) [8].

The WiMedia consists of 256 medium access slots represented by DRP MAS bitmap. After identifying the immediate neighbours, DRP allocation field is set. Now, relay bitmap is initialized with DRP MAS bitmap [6] and the basic procedure of DRP allocation is adopted with hop count as 0. We have proposed the relay allocation procedure which uses the relay allocation format (Fig. 4) consisting of destination (DestinationAddress$_i$) nodes which can be reached by passing through number of nodes (HopCount$_i$) via target node.

Once synchronized, Source node will start receiving relay allocation format through beacon frames. An updation in relay allocation format takes place only if the following constraints are satisfied:

- New destination address is available.
- Existing destination address with higher link quality.
- Sender node is used as NextHop to reach DestinationAddress.

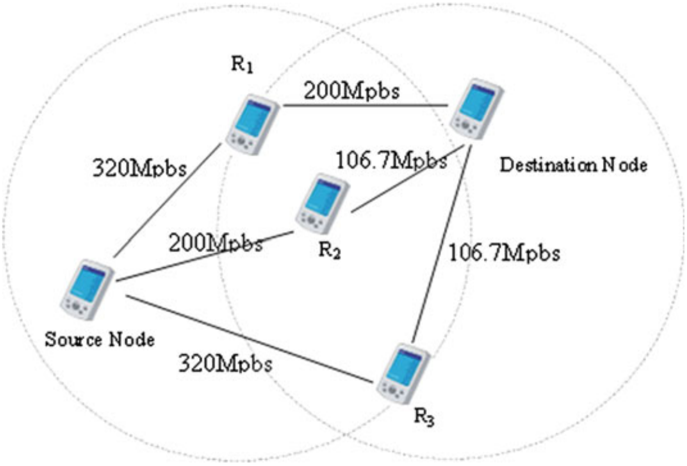
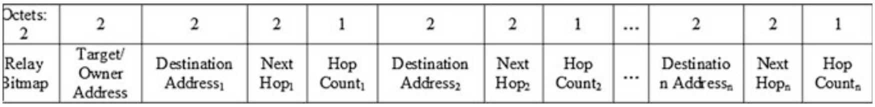


Fig. 3 Basic relay formation using link quality



Octets: 2	2	2	2	1	2	2	1	…	2	2	1
Relay Bitmap	Target/ Owner Address	Destination Address$_1$	Next Hop$_1$	Hop Count$_1$	Destination Address$_2$	Next Hop$_2$	Hop Count$_2$	…	Destination Address$_n$	Next Hop$_n$	Hop Count$_n$

Fig. 4 Relay allocation format

Now, relay allocation format is updated with corresponding destination address having HopCount incremented by 1.

3.1 Proposed Relay Formation Algorithm

Initialise.

```
for all neighbouring nodes do
  set Relay Bitmap equal to DRP Bitmap
  HopCount as 0
  NextHop to Nil
  Target address to Destination Address
endfor
```

Relay Allocation.

```
for every Beacon Frame in the Superframe do
  Access Link Quality Indicator IE and Target Address
  for every Relay Allocation do
    if there is a new target Address and no conflict in
    the slot then
      Add the sender address is target address
      HopCount as HopCount + 1
      Destination Address as Next Hop Address
    else if a matching entry is found with higher Link
    Quality and lesser HopCount then
      Update the sender address is target address
      HopCount as HopCount + 1
      Destination Address as Next Hop Address
    end if
    Forward the updated information to all the neighbours
    using Beacon Frames
  end for
end for
```

Once reservation is completed, the source may start to transmit the data in its identified MAS. During registration process neighbouring nodes identify freely available slots; to handle various slots they prefer store and forward method of multi-hopping. While using incompatible data rates, source node may fragment a packet into multiple chunks limited by mMaxFragmentCount. WiMedia MAC protocol uses minimum interframe space (TMIFS) and short interframe space (TSIFS) in between fragments and frames, if it needs to continue data transfer for more than one slot.

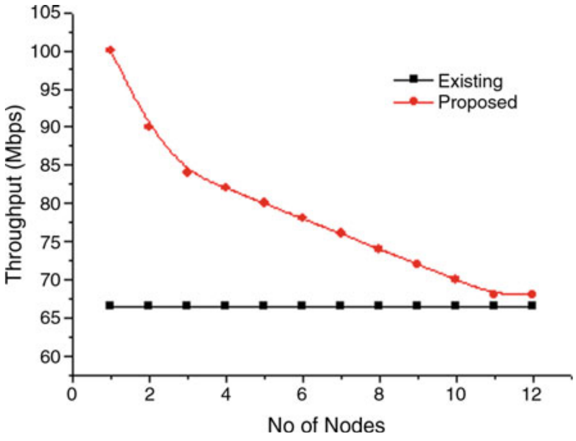
4 Performance Analysis

To analyse the efficiency of the proposed method, OmNet++ tool is used and simulated using inetmanet-2.2 according to WiMedia specifications [10]. We have considered static nodes which are capable of communicating within single hop distance and arranged in linear topology by assuming ideal channel conditions with saturated allocations. The source node generates traffic in consistent manner and the results were observed for varying data rates. The other parameters used for simulation are presented in Table 1.

Table 1 Simulation metrics

Number of node	Varies from 10 to 12
Maximum payload size	4,095 octets
Maximum no. of beacon slots	96 beacon slots
Beacon slot duration	85 µs
Superframe duration	256 × MAS duration
MAS duration	256 µs
Total no. of DRP slots	112 MASs

Fig. 5 Throughput achieved



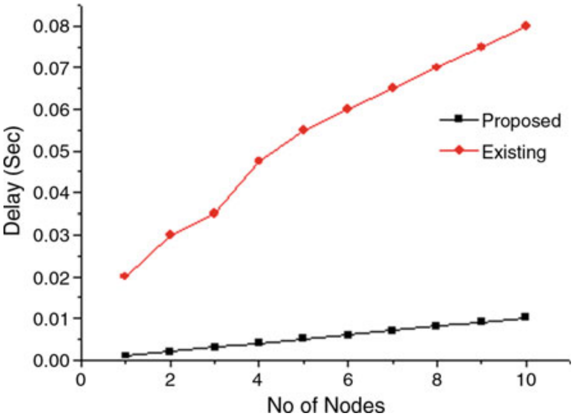


Fig. 6 Effect of number of nodes on delay

The curiosity behind developing multi-hop network is to provide interference free wireless networking. First, we tried to analyse the throughput achieved by increasing the cooperative relays (Fig. 5).

In relay networks, though we are able to maximise throughput when compared to 3-hop allocation procedure [7] (Fig. 5), increase in the number of nodes may influence the throughput. The advantage of doing so is to have a higher data rate. The capacity of the network may be limited due to congestion and control overheads of the participating node.

Another fact is, the change in the network topology is possible due to the selection of relay node. If the distance between source and destination increases or the signal strength is low then it leads to delay. Multi-hop approach attempts to reduce the delay and increase the throughput (Fig. 6). Using the simulation results, we have observed that the proposed method experiences minimal delay for lower number of relays.

5 Conclusions

In this paper, we have proposed a link quality-based multi-hop network for WiMedia Alliance. For establishing the path, pairwise data rate supported is considered. There is a possibility of allocating different slots for single communication channel. In those cases, we propose to store and forward for relaying. Any further updates will reflect on the beacon slot occupancy and needs reconsideration of slot management by adopting the respective changes [11].

We believe that the results of the developed multi-hop network had shown an optimal delay and throughput. The cross-layer design considering cooperative nodes by gathering information from source to the destination node could help fellow researchers to develop further energy-efficient relay networks [12]. The work may further be extended to reliable networks considering the nature of the node, network traffic and present solution for hurdles being faced while finding relays. The proposed algorithm expects commonality in the data rate so that data transfer may experience unique delay and favours bandwidth estimated prior to data transfer. Hence further we address the issued related with various data rate supported by relay nodes.

References

1. Radwan, Ayman, Jonathan Rodriguez (2012). Energy saving in multi-standard mobile terminals through short-range cooperation. EURASIP Journal on Wireless Communications and Networking , 1–15. doi:10. 1186/1687-1499-2012-159
2. Yang, Zhuo, et al (2010). A TDMA-based MAC protocol with cooperative diversity. IEEE communications letters, 542–544. doi:10.1109/LCOMM.2010.06.092451
3. Lee, Jong-Kwan, Hong-Jun Noh, Jaesung Lim(2012). Dynamic cooperative retransmission scheme for TDMA systems. IEEE Communications Letters, Vol 16, Is 12, 2000–2003. doi:10.1109/LCOMM.2012.101712.121854
4. Zhou, et al(2011). Link-utility-based cooperative MAC protocol for wireless multi-hop networks. IEEE Transactions on Wireless Communications, vol 10, Is 3, 995–1005. doi:10. 1109/TWC.2011.122010.100773
5. Lee, Jong-Kwan, Hong-Jun Noh, Jaesung Lim (2014). TDMA-based cooperative MAC protocol for multi-hop relaying networks. IEEE Communications Letters, Vol 18, Is 3, 435–438. doi:10.1109/LCOMM.2014.011314.132095
6. Specification for Distributed MAC for Wireless Networks version 1.5, WiMedia Alliance, 2009.
7. Park, Seunghyun, Hyunhee Park, Eui-Jik Kim (2014). Distributed Relay-Assisted Retransmission Scheme for Wireless Home Networks. International Journal of Distributed Sensor Networks 2014. doi:http://dx.doi.org/10.1155/2014/683146
8. Kim, Junwhan, Jaedoo Huh (2007). Rate adaptation scheme for slot reservation in WiMedia MAC. Consumer Electronics, 2007. ICCE 2007. Digest of Technical Papers. International Conference on. IEEE, 2007. doi:10.1109/ICCE.2007.341403
9. Joo, Yang-Ick, Kyeong Hur (2011). A multi-hop resource reservation scheme for seamless real-time QoS guarantee in wimedia distributed mac protocol. Wireless Personal Communications ,Vol 60, No. 4, 583-597. doi:10.1007/s11277-010-9961-3
10. Alam, Muhammad, et al(2013). Energy and Throughput Analysis of Reservation Protocols of Wi Media MAC. Journal of Green Engineering, 363–382. doi:10.13052/jge1904-4720.341
11. Vishnevsky, V.M.; Lyakhov, A.I.; Safonov, A.A.; Mo, S.S.; Gelman, A.D., “Study of Beaconing in Multihop Wireless PAN with Distributed Control,” Mobile Computing, IEEE Transactions on, vol.7, no.1, pp.113, 126, Jan. 2008 doi:10.1109/TMC.2007.1078.
12. Li, Fulu, Kui Wu, Andrew Lippman (2006). Energy-efficient cooperative routing in multi-hop wireless ad hoc networks. Performance, Computing, and Communications Conference, 2006. IPCCC 2006. 25th IEEE International. IEEE, 2006. doi:10.1109/.2006.1629410

Assessment of Reusability Levels on Domain-Specific Components Using Heuristic Function

N. Md. Jubair Basha and Sankhayan Choudhury

Abstract Process reuse has noticeable role in the software component reuse, increasing the prominence in enterprise software development. The research gap identified from the related work is to identify reusable components from the legacy system. In order to fill this research gap, a methodology to assess the reusability of components has been proposed using a heuristic function. The proposed methodology is realized and implemented using three domain applications for the assessment of reusability levels. The heuristic function hampers the non-reusable components from the assessed reusability levels and helps to identify the reusable components from the legacy systems.

Keywords Domain-specific components · Reusability levels · Decision-to-decision path · Independent path · Reachability matrix · Heuristic function

1 Introduction

In software industry, the concept of software reuse has existed since the beginning of programming, as programmers reuse algorithms, sub-routines, and segments of code from previously created programs. The idea of reuse in software was first formalized by Mcllory [1], who stressed the need to componentize software systems. Mcllory's ideas were directed to thoughts about building software systems in a similar manner to building hardware systems (for example, electronic circuits). Hence, more advanced research work emerged that discussed reuse and its possible

N.Md. Jubair Basha (✉)
IT Department, Muffakham Jah College of Engineering & Technology,
Hyderabad, India
e-mail: jbasha@acm.org

S. Choudhury
CSED, University of Calcutta, Kolkata, India
e-mail: sankhayan@gmail.com

H.S. Saini et al. (eds.), *Innovations in Computer Science and Engineering*,
Advances in Intelligent Systems and Computing 413,
DOI 10.1007/978-981-10-0419-3_19

directions, emphasizing the significance and need for reuse [2, 3]. Nowadays, reuse has become one of the standard paradigms that most leading software development vendors such as IBM, HP, and Motorola practice in their production lines and many others have reported eminent experiences with applying reuse in their software development projects [4].

Rather than developing from scratch, identifying the reusable components from the legacy systems will be an advantage to reduce cost, effort, and time to the developers. The developer needs to identify reusable components from the legacy systems. So, a methodology has been proposed to assess the reusable components using heuristic function. This paper is organized as follows. In Sect. 2, the related work with detailed literature pertaining to reusability assessment is discussed. Section 3 discusses the need of the proposed work and exposes the drawbacks of the previous work. This section provides the motivation to this paper. In Sect. 4, the proposed methodology is realized and implemented using three domain applications and discusses the need of heuristic function and applies it on the proposed work. Section 5 concludes the paper.

2 Related Work

Software reuse is the utilization of available software or to build new software from software knowledge. Reusable assets can be any reusable software or software knowledge. Reusability is a property of a software asset that indicates its probability of reuse [5]. Software reuse means the process that uses "designed software for reuse" again and again [6]. By reusing software, it is possible to manage complexity of software development, increase product quality and make production faster in the organization.

Software reuse can be broadly classified into two categories, i.e., product and process reuse. Product reuse involves the reuse of a software bit and producing a new component as an outcome of module integration and construction. Process reuse represents the reuse of legacy components from repository. These components may be directly reused or may need minor improvements. The improved software component can be archived by converting these components. The components may be classified and chosen depending on the required domain [7]. It can be improved by identifying objects and operations for a class of equivalent systems, i.e., for a specific domain. In the context of software engineering, domains are application areas [8].

As a part of process reuse the relevant work is identified with the respective related work. Sharma et al. [9] proposed a neural network based approach to predict reusability of a software component. The work considered only four attributes such as customizability, portability, complexity of interface and understandability, which influence the reusability of black-box components. These four attributes are considered as input parameters and reusability is the input output parameter to train the network. Both training and testing are performed by a different number of hidden layers and neurons to get the best results. These results show that the network is able to predict the reusability of the components with accepted precision. The drawback identified in this work is the level of reusability considered for only particular attributes, but it still lacks with the reuse of white box of the components. Sagar et al. [10] extend the work in [9] to estimate the reusability of software bits by using Mamdani-based fuzzy inference system. This work is verified against two small classroom-based components. However, it lacks more rigorous validation on complex commercial real life applications. Singh et al. [11] suggested a soft computing technique to automatically predict software reusability levels i.e., very low, low, medium, high and very high. The neuro-fuzzy approach with the data sets generated by the fuzzy logic is to take advantage of some of the useful features of a neuro-fuzzy approaches such as learning ability and good interpretability. The reusability measures were predicted only based on the performance but not with the behavioral properties of the system.

Gandhi et al. [12] proposed metrics measure quantitative generic construct with inheritance in an object-oriented code. Two metrics are proposed, namely GRr (generic reusability ratio) and ERr (effort ratio). First metric GRr estimates impact of template in program volume and second metric ERr measures impact of template in evolution effort. These metrics act as tools for estimating and evaluating costs of program design and program tests as well as program complexity. Mohr [13] presented a formally described vision statement for the estimation of practical reusability of services and sketches an exceptional reusability metric that is based on the service descriptions. Services are self-contained software bits that can be used as platform independent and that aim at maximizing software reuse. A primary concern in service-oriented architectures is to measure the reusability of services. The metrics for functional reusability of software either require source code analysis or have very little explanatory power. Here, the research gap is identified to consider source code analysis using a heuristic function to assess the reusability of domain-specific components. With this, it is easy to identify the domain-specific reusable components.

3 Motivation

As per customer requirements, the developer needs to identify the common behavior of the components from the legacy components. The common behavior of the components is reusable. The dire need to assess the reusability levels from the literature review [14] of the works in the area of reusability assessment is conducted and the results of the review are presented. Many of the approaches are based on metrics and are applicable to the object-oriented paradigm; the target language used for the application is Java. Only a few studies have used experimentation to validate. This needs the attention of the current research community to gain confidence of the software practitioners. Further there is a dire need to explore the work on domain specificity of the components. The research gap identified from the above related work motivates to propose a methodology for the assessment of reusability levels on the domain-specific components from the legacy systems by using a heuristic function.

4 A Heuristic Function Based Methodology to Assess the Reusability Levels of Domain Specific Components

The motivation clearly shows the research gap identified from the literature. The proposed methodology will solve the problem identified as follows:

1. Step I: Consider the already available applications from the repository.
2. Step II: Identify the different components from the available legacy applications.
3. Step III: Consider the source line of code (SLOC) of each existing component.
4. Step IV: Identify the independent paths (CI_M) from the decision-to-decision (DD) graph.
5. Step V: Derive the reachability matrix (RM_{CL}) from the DD graph for every component.
6. Step VI: Repeat Step V for each component of different applications.
7. Step VII: Check how many times each path is traversed using the reachability matrix (RM_{CL}) from the DD path for each component.
8. Step VIII: Finally, check with Heuristic Function (H_{Reuse}) = $C_{RL} > RM_{CL}$ for assessing the reusability levels of domain-specific components.
   Where $RM_{CL} = 2$, i.e., independent paths of each component derived from reachability matrix and C_{RL}= reusability level of component.

4.1 Decision-to-Decision Path

A decision-to-decision (DD) path, is a path of execution (usually through a flow graph representing a program, such as a flow chart) between two decisions. Current versions of the concept also include the decisions themselves in their own DD-paths.

The proposed methodology is realized with the following domain-specific components. The DD path in Fig. 1, the temperature converter system is considered with the connection nodes and different links among them, through which it is able to find the independent paths in the respective code or the program. In Fig. 1, the DD graph of converter component is generated and this graph contains A to L nodes. Each node represents the flow of execution in the program. For the above DD path since the cyclomatic complexity is 4, therefore we can create four independent paths for this component. They are A-B-C-D-E-L, A-B-C-D-E-F-G, A-B-C-D-E-F-H, A-B-C-D-E-F-I.

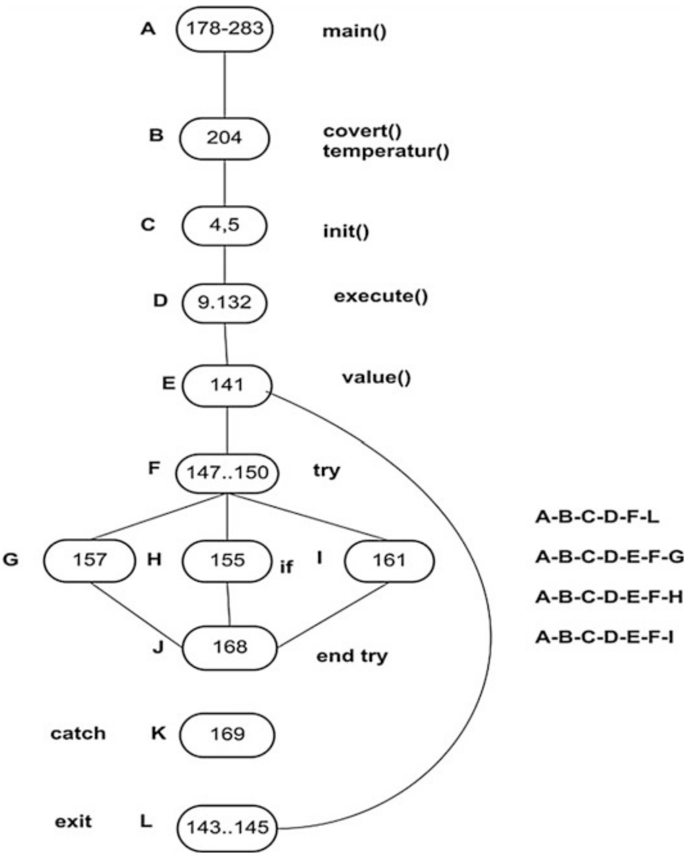


Fig. 1 DD path for converter component of temperature converter system

The DD path for result component of student management system is presented in Fig. 2. The connection of nodes and the different links among them is to identify the independent paths with the respective code or the program. The graph represented with the matrix can be named as reachability matrix (RM_{CL}). Each independent path is represented as ‘1’ in the matrix between two or more nodes named as an independent component path (C_{IM}). The matrix value with 1 can be denoted as the reused component. This is the traced path among the components for a particular behavior or functionality. In Fig. 2, the DD path of the result component is represented with four nodes and these nodes are designated as A, B, C, D, respectively. The two paths from A to D, i.e., A-B-D and A-C-D, the independent paths (C_{IM}) are identified and simultaneously the matrix with these paths can be considered.

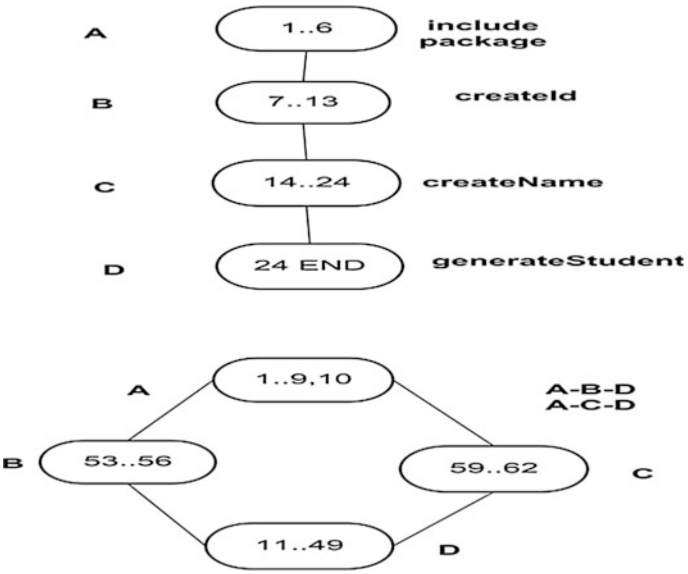


Fig. 2 DD path for result component of student management system

4.2 Reachability Matrix (RM_{CL})

Definition:

Let D = (V, A) be a digraph, where V = (1, 2, 3.....n). The adjacency matrix of the digraph D is n x n matrix, A where a_{ij} the entry on the ith row and jth coloumn, is defined by a_{ij} = 1 if (i, j) belongs to A or 0 if (i,j) doesnot belongs to A.

The reusability matrix (RM_{CL}) of the digraph D is an n x n matrix R where r_{ij} the entry on ith row and jth coloumn is defined by r_{ij} = 1 if j is reachable from i or 0 if j is not reachable from j.

Table 1 represents the reachability matrix (RM_{CL}) for the result component of the student management system. In this graph it is found that there are two independent paths, i.e., A-B-D and A-C-D. In the path A-B-D the matrix for A-B is represented as 1 similarly B-D is taken as 1. So the path A-B-D is an independent path and it can be reused. Table 2 represents reachability matrix (RM_{CL}) for ApplicationEntity component of the respective DD path. From this, it is found that there were no independent paths. So the matrix value for this component is

Table 1 Result component

	A	B	C	D
A	0	1	1	1
B	1	0	0	0
C	1	0	0	0
D	1	0	0	0

Table 2 ApplicationEntity component

	A	B	C	D
A	0	0	0	0
B	0	0	0	0
C	0	0	0	0
D	0	0	0	0

Table 3 Converter component of temperature converter system

	A	B	C	D	E	F	G	H	I	J	K	L
A	0	1	1	1	1	0	1	1	1	0	0	1
B	1	0	1	0	0	0	0	0	0	0	0	0
C	1	1	0	1	1	0	0	0	0	0	0	0
D	1	0	0	0	1	0	0	0	0	0	0	0
E	1	0	0	0	1	1	0	0	0	0	0	0
F	0	0	0	0	0	1	1	0	0	0	0	0
G	1	0	0	0	0	0	0	0	0	0	0	0
H	1	0	0	0	0	0	0	0	0	0	0	0
I	1	0	0	0	0	0	0	0	0	0	0	0
J	0	0	0	0	0	0	0	0	0	0	0	0
K	0	0	0	0	0	0	0	0	0	0	0	0
L	1	0	0	0	0	0	0	0	0	0	0	0

represented as 0. Since there are no traceable paths, it is identified that this component cannot be reused.

Table 3 represents the reachability matrix (RM_{CL}) for converting component drawn for the respective DD path. It is identified that there are four independent paths, i.e., A-B-C-D-E-L, A-B-C-D-E-F-G, A-B-C-D-E-F-H, and A-B-C-D-E-F-I. From the path, A-B-D the matrix for A-B is represented as 1, B-C = 1. C-D = 1, D-E = 1, and E-L = 1. The path A-B-C-D-E-L is an independent path which can be reused. Similarly in the reachability matrix (RM_{CL}), the independent paths A-B-C-D-E-F-G, A-B-C-D-E-F-H, A-B-C-D-E-F-I are identified and the respective value of 1 is taken for these paths. Table 4 represents the independent paths of the respective application; each independent path possesses the series of executable lines of code which are compiled, irrespective of the other part of it. Hence those

Table 4 Independent paths of the different domain applications

Application name	Independent path (CI_M)
Temperature converter system	A-B-C-D-E-L, A-B-C-D-E-F-G, A-B-C-D-E-F-H, A-B-C-D-E-F-I
Student management system	A-B-D, A-B-C, A-C-D
Library management system	A-B-D, A-C-D

Table 5 Level of reuse of different domain-specific components

Component name	Level of reuse (C_{RL})
ApplicationEntity	0
CreateServlet	7
EntityFacade	7
Entity façade remote	9
Convert	2
Converter	6
ServiceApprehensive	8
Invert	5
Credit	1

independent paths denote the method or the lines of code executed independent of others. The reusability level is determined from the number of traced paths during the execution by considering the independent paths (C_{IM}). Table 5 represents the reusability level (C_{RL}) of each component identified from the independent paths and generated from the reachability matrix (RM_{CL}).

The need of heuristics is considered because for domain specificity, the components' behavior may vary from one domain to another. Usually, heuristics will be applicable when there is a decision making on a particular condition-based event. So, the heuristic function (H_{Reuse}) restricts up to some level which is a desired value required by any application domain for the possibility of reuse. The heuristic function aids to assess the reusability level of domain-specific components for the considered multiple domains.

$$\text{The Heuristic Function } (H_{Reuse}) = C_{RL} > 2$$

where $C_{RL} \geq 2$, i.e., independent paths of each component derived from reachability matrix (RM_{CL}).

The heuristic function (H_{Reuse}) may vary from domain to domain. By applying heuristic function (H_{Reuse}), as there must be minimum two independent paths from the reachability matrix (RM_{CL}). If the level of reuse is less than 2, then the components cannot be reused for the respective domain and there must be minimum two independent paths. If the reusability level (C_{RL}) is greater than 2, only then the components can be applicable. Accordingly, from the traced path, the assessment of reuse level (C_{RL}) is applied on domain-specific components using heuristic function.

5 Conclusion

The concept of reuse greatly benefits the software industry. The related work provides research gaps in the assessment of reusability levels. This motivates to propose a methodology for the assessment of reusability levels using heuristic function. The proposed methodology is realized and implemented on the components of three domain applications. The advent of heuristic function (H_{Reuse}) restricts the non-reused components and avoids such components for the identification of reusable components from the legacy system. As a part of future work, the domain-specific component interactions can be identified and different heuristics can be applied to generate reusable components.

References

1. McIlroy, M. Douglas, et al. "Mass-produced software components." Proceedings of the 1st International Conference on Software Engineering, Garmisch Pattenkirchen, Germany. sn, 1968.
2. Mili, Hafedh, Fatma Mili, and Ali Mili. "Reusing software: Issues and research directions." Software Engineering, IEEE Transactions on 21.6 (1995): 528–562.
3. Frakes, William B., and Kyo Kang. "Software reuse research: Status and future " IEEE transactions on Software Engineering 31.7 (2005): 529–536.
4. Almeida, Eduardo S., et al. "CRUISE: Component Reuse in Software Engineering." (2007).
5. Zhongjie Wang et.al, "A Survey of Business Component Methods and Related Technique", World Academy of Science, Engineering and Technology, pp.191–200, 2006.
6. N Md Jubair Basha, Salman Abdul Moiz, "A Methodology to manage victim components using CBO measure", published in International Journal of Software Engineering & Applications(IJSEA), Vol3 (2) pp 87–96, March 2012, ISSN:0975-9018.
7. Jianli He, Rong Chen, Weinan Gu: A New Method for Component Reuse, IEEE Transactions on Software Engineering, 2009.
8. Maurizio Pighin: A New Methodology for Component Reuse and Maintenance, IEEE Transactions on Softwrae Engineering, 2001.
9. Sharma, A., Grover, P. S., & Kumar, R. (2009). Reusability assessment for software components. *ACM SIGSOFT Software Engineering Notes*, *34*(2), 1–6.
10. Sagar, S., Nerurkar, N. W., & Sharma, A. (2010). A soft computing based approach to estimate reusability of software components. *ACM SIGSOFT Software Engineering Notes 35* (5), 1–5.
11. Singh, Y., Bhatia, P. K., & Sangwan, O. (2011). Software reusability assessment using soft computing techniques. *ACM SIGSOFT Software Engineering Notes*, *36*(1), 1–7.
12. Gandhi, P., & Bhatia, P. K. (2011). Estimation of generic reusability for object-oriented software an empirical approach. *ACM SIGSOFT Software Engineering Notes*, *36*(3), 1–4.
13. Mohr, F. (2014). A Metric for Functional Reusability of Services. In *Software Reuse for Dynamic Systems in the Cloud and Beyond* (pp. 298–313) Springer.
14. Fazal-e-Amin, A. K. M., & Oxley, A. (2011). A Review of Software Component Reusability Assessment Approaches. *Research Journal of Information Technology*, *3*(1), 1–11.

Reliable Propagation of Real-Time Traffic Conditions in VANETS to Evade Broadcast Storm

Mohd Umar Farooq, Mohammad Pasha and Khaleel Ur Rahman Khan

Abstract Vehicular communication poses a challenge of high mobility of vehicles and the preeminent solution for communication is to broadcast. But it leads to packet redundancy followed by broadcast storm and congestion. This paper proposes a steering strategy for the constant movement conditions utilizing vehicular ad hoc network. The noticeable feature of the proposed paper is that it minimizes Broadcast—Storm—Problem (BSP). The alert message (AM) is transmitted among the road side units (RSU) while the vehicles on street are only in transmitting mode to RSUs. The victimized vehicle creates an alert message that is transmitted to the closest RSU. The message is then sent to all the RSUs on that path utilizing I2I (infrastructure to infrastructure) correspondence. The RSUs process the alert message and signal the vehicles using light emitting diodes (LEDs) planted along the path. Our proposed system decreases Packet—Lost—Ratio (PLR), redundancy, or duplication of messages from numerous foundations. Congestion controlling is enhanced by the fact that the message need not be rebroadcasted by vehicles. Our work is suitable for generic situations.

Keywords MANET · VANET · Broadcast storm problem · Victimized vehicle · Light emitting diodes · Road side units

M.U. Farooq (✉)
Department of CSE, MJCET, Osmania University, Hyderabad, India
e-mail: umarfarooq.mohd@gmail.com

M. Pasha
Department of IT, MJCET, Osmania University, Hyderabad, India
e-mail: muhammed.pasha@gmail.com

K. Ur Rahman Khan
Department of CSE, Ace College of Engineering, Hyderabad, India
e-mail: khaleelrkhan@gmail.com

H.S. Saini et al. (eds.), *Innovations in Computer Science and Engineering*,
Advances in Intelligent Systems and Computing 413,
DOI 10.1007/978-981-10-0419-3_20

1 Introduction

Presently traffic congestion has turned into a real issue of concern. At first vehicle route framework was supported by computerized maps incorporated with GPS recipients. Then again, in this method there are no real-time traffic conditions. This setback emerges the requirement for identifying continuous movement conditions of the nodes to overcome traffic-related issues. Vehicular Ad hoc Networks (VANETS) was introduced to address these traffic issues. It is a remote system which is established between vehicles and road side units on a street to impart some critical bit of data which can help decrease road blockages. This paper proposes a directing method for ongoing movement situations considering the dynamic movement of traffic. The road side units (RSU) entity is used for actualizing this thought. At whatever point a traffic issue happens that may cause an aggravation in the flow of movement, an alert message is transmitted to the closest RSU. RSUs store the alert message for some amount of time, process it, and forward it to the next RSU. They do not generally contain the same message. At the point when a vehicle enters a street with a road block and when it enters the range of an RSU of that street, it gets a signal from it informing about some issue ahead by LED lighting and the vehicles can be re-routed. The scenario after the traffic clearance ought to be presented likewise. Envision that after a mishap the path is going to be cleared in next 1 min. At that point it is clear that the approaching traffic ought to keep proceeding onward the same path as it will be cleared soon for smooth movement of vehicles. This might be kept up by utilizing LEDs planted in the road. The LEDs turn orange-red in case of an impediment ahead.

2 Related Work

Survey has shown that disseminating real-time traffic information is in the form of reports which are prioritized in terms of their value, as reflected by supply and demand [1]. Each vehicle makes a local decision on when to disseminate a report in order to deal with the bandwidth and memory constraints. The study of dissemination of real-time traffic conditions is carried out by dividing vehicles into clusters and assigning header and trailer to efficiently disseminate information warning functions [2]. Acknowledgement-based broadcast protocol only employs local information acquired via periodic beacon messages, containing acknowledgements of circulated broadcast messages [3]. Distributing messages among highly mobile hosts using direct radio communication between moving vehicles on the road that require no additional infrastructure [4]. Sending messages to mobile users in disconnected ad hoc wireless network in which mobile hosts actively modify their trajectories to transmit messages [5]. The problem can be also defined as “The loss due to excessive amount of redundant traffic, exaggerated interference among neighboring nodes and limited coverage" [6]. The vehicles are compacted on a road and the average number of vehicles that occupy one mile of road is beyond a definite value that results in congestion and redundant in broadcast of messages among neighboring nodes (the problem we refer to as broadcast storm problem) [7]. This method uses a TLO algorithm to solve the broadcast storm problem. In this method the victimized vehicle directs the message to all the vehicles next to it. The vehicles which receive the message do not rebroadcast instantly but run a TLO algorithm to detect the last vehicle. But the vehicles which receive the message do not rebroadcast it immediately, instead run a TLO algorithm to detect the last vehicle.

3 Proposed System

In our proposed framework, each vehicle needs to be equipped with only a transmitter. Vehicles don't transmit messages to other vehicles; they can transmit messages to RSUs. They transmit a message to the closest RSU when they run over a moderate moving traffic or a barricade. The point of interest of not permitting vehicles to transmit messages to different vehicles is that repetition will be disposed-off to a considerable degree. There will not be postponement in any packet in light of the fact that alert message is indicated using LEDs. Unnecessary flow of messages can be avoided because if vehicles can transmit messages; then message propagation will be infinite resulting in congestion. A vehicle may receive many alert messages from various affected zones and may lead to disturbance to the vehicle. To avoid these overheads, this paper proposes a framework where the vehicles transmit message only to road side units (RSU) which in turn transmit messages to other RSUs. The status of the traffic is known by LEDs implanted in the road. Synchronization plays a major role in packet dissemination in this system where RSUs do not just keep on receiving alert packets but they stop accepting the packets after getting the first alert message. It cannot process any message until the lock on RSU is released. However, the RSUs are in the listening mode for a small period of additional time to gauge the density of vehicles based on the alert messages received. After receiving an alert message, the RSU goes into the blocked state and does not accept messages from other vehicles. However, the number of requests sent to the RSU in listening mode is stored, say '*r*', is used to calculate the density of traffic and an estimated time to clear the traffic, say '*t*', is calculated. RSU remains in the blocked state, say '*t*' seconds and then goes back to the unblocked state. If an alert message is immediately received by the same RSU, it goes into the blocked state again and the same procedure is continued. Taking a situation and chipping away a little activity is insufficient.

Consider a situation where many vehicles move in a path. In such a situation, if the vehicle gets various measures of unnecessary messages, it would result in a ton of unsettling influence. A vehicle might rather want to get an alert message relating to the way the vehicle is proceeding onward. The figure beneath demonstrates the usefulness of our proposed framework and the way it suits best for constant movement conditions. At whatever point a sensor distinguishes a barrier, it sends an alert message to the closest RSU and after that the RSU advances that message to the various RSUs on that street. The vehicles approaching the path recognize the orange-red light from the LEDs with the goal that they can take a passageway and pick an interchange course. It is proposed that in the case of a road block, the LEDs turn orange-red and in clear traffic conditions, the LEDs are turned off. To keep the RSUs from transmitting messages along the streets of long length, the idea of hop count is presented. As per this, in the situation of accepting an alert message from a vehicle, RSU can forward this message to a particular number of RSUs. This way, the vehicles on the same path of the road and far from the road block are not bothered. Improvement of packet dissemination is accomplished by avoiding rebroadcast (Fig. 1).

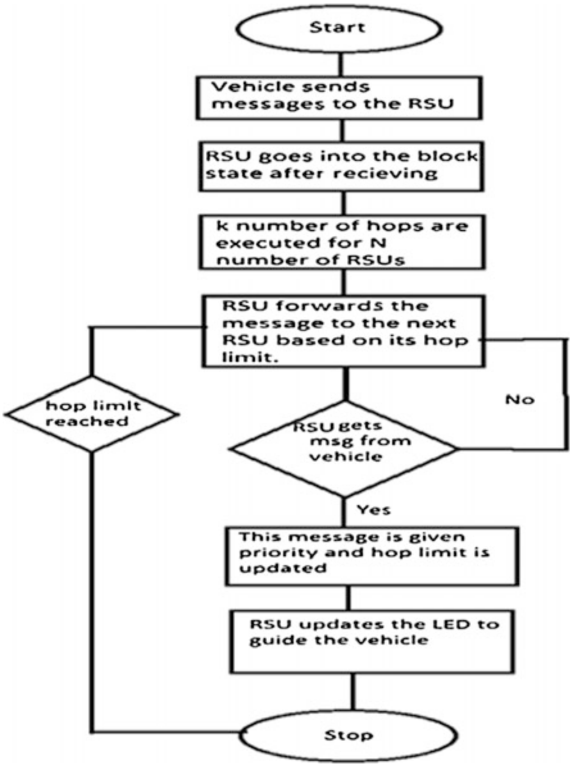

Start
Vehicle sends messages to the RSU
RSU goes into the block state after recieving
k number of hops are executed for N number of RSUs
RSU forwards the message to the next RSU based on its hop limit.
hop limlt reached
No
RSU gets msg from vehicle
Yes
This message is given priority and hop limit is updated
RSU updates the LED to guide the vehicle
Stop

Fig. 1 System workflow for processing alert messages by RSUs

3.1 Proposed System Workflow

Algorithm

1. Start.
2. The Vehicles send *TRAFFIC INFO* message to RSU (say ‘*R1*’).
3. *R1* goes into the blocked state after receiving the message from a vehicle.
4. The message is forwarded to other ‘*N*’ RSUs based on its hop limit (say ‘*k*’ hops).
5. If an RSU receives a *TRAFFIC INFO* message directly from a vehicle, it is given priority over the message received from another RSU with originator as the same vehicle. Hop limit is updated.
6. The RSUs update the LEDs to guide the Vehicles.
7. Based on the LED status, re-routing decision is made at the nearest junction.

3.2 Proposed Model

The following parameters were considered while demonstrating the working of the proposed algorithm.

- Number of RSUs ‘*N*’.
- Number of Packets transmitted ‘*p*’.
- Hop limit of RSUs for forwarding the Alert Messages ‘*k*’.
- Number of unique requests or Number of Vehicles ‘*n*’.
- Estimated LED status time conveying Road Blockage or Clearance ‘*T*’.
- Interval of receiving *TRAFFIC INFO* Messages from Vehicles ‘Δt’.
- Priority for *TRAFFIC INFO* ‘*P*’.

The above proposed system signifies that *TRAFFIC INFO* Messages are disseminated by vehicles only destined to RSUs resulting in efficient traffic flow and reduced network congestion.

3.3 Analysis and Results

The proposed idea is implemented in OMNET++. The number of redundant broadcasts is reduced to a large extent as shown below. The metrics used to evaluate our proposed system are delay and throughput. Estimated LED status time conveying road blockage or clearance ‘*T*’ is calculated with respect to the vehicle

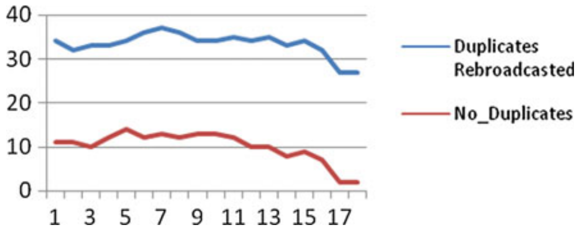


Fig. 2 Number of broadcasted alert messages

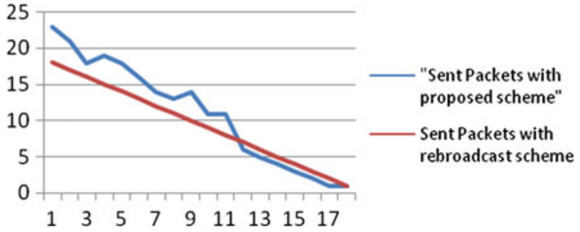


Fig. 3 Sent packets comparison

density. The data packets that are successfully delivered to the destination are counted. Mathematically, it is calculated as (Figs. 2 and 3),

$$\sum(\text{arrive time} - \text{send time})/\sum \text{Number of connections}$$

4 Conclusion

In this paper we introduced a scheme which establishes an effective and reliable communication between vehicles and road side units (RSUs). The key idea behind the proposed scheme is to control the number of packets in the network by making vehicles only transmit to RSUs and avoid rebroadcast among vehicles. The vehicle re-route decision is assisted by the LED status planted along the road.

References

1. Li and Daniela Rus, C. Hu, Y. Hong, and J. Hou, "On mitigating the broadcast storm problem with directional antennas," in *Proc. IEEE International Conf. On Commun. (ICC)*, vol. 1, Seattle, USA, May 2003, pp. 104–110.
2. S. Ni, Y. Tseng, Y. Chen, and J. Sheu, "The broadcast storm problem in a mobile ad hoc network," in *Proc. ACM Intern. Conf. on Mobile Computing and Networking (MOBICOM),* Seattle, USA, 1999, pp. 151–162.

3. Kanitsorn Suriyapaibonwattan and ChotipatPornavalai, "An Effective Safety Alert Broadcast Algorithm for VANET", in *International Symposium on Communication and Information Technology*, 2008.
4. Ashwin Gumaste and Anirudha Sahoo VEHACOL:VehicularAntiCollisionMechanism using a Combination of PeriodicInformation Exchange and Power Measurements.
5. A Stable Routing Protocol to Support ITS Services in VANET Networks.
6. M. Shulman and R. Deering, "Third annual report of the crash avoidance metrics partnership April 2003–March 2004," Nat. Highw. Traffic Safety Admin. (NHTSA), Washington, DC, Jan. 2005. DOT HS 809 837.
7. Ting Zhong, Bo Xu, Piotr Szczurek, Ouri trafficinfo: an algorithm for vanet dissemination of real-time traffic information1.

Hybrid Multi-objective Optimization Approach for Neural Network Classification Using Local Search

Seema Mane, Shilpa Sonawani and Sachin Sakhare

Abstract Classification is inherently multi-objective problem. It is one of the most important tasks of data mining to extract important patterns from large volume of data. Traditionally, either only one objective is considered or the multiple objectives are accumulated to one objective function. In the last decade, Pareto-based multi-objective optimization approach have gained increasing popularity due to the use of multi-objective optimization using evolutionary algorithms and population-based search methods. Multi-objective optimization approaches are more powerful than traditional single-objective methods as it addresses various topics of data mining such as classification, clustering, feature selection, ensemble learning, etc. This paper proposes improved approach of non-dominated sorting algorithm II (NSGA II) for classification using neural network model by augmenting with local search. It tries to enhance two conflicting objectives of classifier: Accuracy and mean squared error. NSGA II is improved by augmenting backpropagation as a local search method to deal with the disadvantage of genetic algorithm, i.e. slow convergence to best solutions. By using backpropagation we are able to speed up the convergence. This approach is applied in various classification problems obtained from UCI repository. The neural network modes obtained shows high accuracy and low mean squared error.

Keywords Classification · Neural network · Multi-objective optimization · Pareto optimality · NSGA II · Local search

S. Mane (✉) · S. Sonawani
Maharashtra Institute of Technology, Pune, India
e-mail: seemamane491@gmail.com

S. Sonawani
e-mail: shilpa.sonawani@mitpune.edu.in

S. Sakhare
Vishwakarma Institute of Information Technology, Pune, India
e-mail: sakharesachin7@gmail.com

H.S. Saini et al. (eds.), *Innovations in Computer Science and Engineering*,
Advances in Intelligent Systems and Computing 413,
DOI 10.1007/978-981-10-0419-3_21

1 Introduction

Data mining is nothing but KDD, i.e. knowledge discovery in large databases. Discover the important patterns and knowledge from huge data which are useful in decision-making systems. So, the aim of data mining process is to build optimal predictive or descriptive model which best fit data and by using these models we can extract knowledge and patterns from large databases. Classification is one of the most frequently used tasks for decision-making in human activity. We can use classification when we need to allocate predefined class or group to unlabelled object by observing various attributes of that object and other labelled objects. Classification is inherently multi-objective problem. Standard mathematical techniques cannot solve classification problem because of complexity. In the last decade, multi-objective approaches are used to solve classification problem efficiently. In literature, different evolutionary algorithms are used to solve multi-objective problems like genetic algorithm, genetic programming, evolutionary algorithm, etc. [1, 2]. Traditionally, different approaches are used to solve multi-objective optimization problems like weighted approach, lexicographic approach and Pareto approach. In weighted approach, multi-objective problem is first converted into single-objective problem and then solve it using single-objective optimization approach. But there are certain real-life problems having more conflicting objectives, which need to be optimized simultaneously to obtain the set of optimal solutions. Classification is multi-objective problem and different objectives for classification problem can be as accuracy, sensitivity, mean squared error, precision, specificity, etc. Therefore, the multi-objective optimization approach is highly applicable to classification problems. Single optimal solution is generated in final generation by single-objective optimizer, while in multi-objective optimization, set of Pareto optimal solutions are generated in final generation, where one objective performance can be increased only by degrading the performance of at least one or more other objectives. Multi-objective evolutionary algorithms (MOEAs) [3] have become popular in data mining. MOEAs can be used for solving data mining task such as classification, regression, clustering, ensemble learning.

The paper is organized as, Literature review is given in Sect. 2 under heading of Related Work. Background knowledge is provided in Sect. 3. Section 4 gives architecture for Hybrid NSGA II. Dataset description and Results are given in Sect. 5. Conclusion and future scope are subsequently drawn in Sect. 6.

2 Related Work

In [4], authors proposed multi-objective optimization approach for financial forecasting using artificial neural network [4]. Evolutionary approach was used to make predictions of the progress of stock market based on NEAT (Neuro Evolution of

Augmenting Topologies [5]). Greedy mutation operator is used for automatic adjustment of weight parameters of current neural network.

In [6], for breast cancer diagnosis author proposed multi-objective approach based on Pareto differential evolution (PDE) [7]. To overcome the disadvantage of evolutionary algorithm, local search technique, i.e. backpropagation is used as to speed up the convergence [6].

Ibrahim et al. used NSGA II [8] for optimization of Three-Term Backpropagation Neural Network [9]. The NSGA II [8] is applied to optimize structure of TTBPN [9] by simultaneously reducing complexity in terms of hidden nodes and error.

In paper [10], multi-objective optimization approach has used system identification problem using recurrent neural network. For encoding of chromosomes variable-length representation is used. With the help of structural mutation neural network architecture is evolved [10]. Microgenetic algorithm is used to optimize multi-objective evolutionary recurrent neural network. Connection weights and network architecture was evolved simultaneously [10].

Renata Furtuna, Silvia Curteanu and Florin Leon proposed multi-objective optimization approach for polysiloxane synthesis [11]. A feed-forward neural network was used with NSGA II [8]. Elitism is used to preserve best individuals in the current generation which are used for next generation as there may be chance of getting loss of good individuals due to crossover and mutation operators.

In paper [12], authors used multi-objective approach for prediction of patient survival after transplantation of liver using evolutionary artificial neural networks [12]. A radial basis function neural network was trained using NSGA II [8]. The neural network models from Pareto fronts were used to build rule-based system which was used to find accurate donor-recipient match [12].

Gossard et al. proposed multi-objective optimization strategy to optimize the envelope of a residential building based on its thermal performance [13]. They have used artificial neural network and NSGA II. Two objectives have been considered: the annual energy load Q_{TOT} and the summer comfort index I_{SUM}. The role of ANN is to provide fast and accurate evaluations of objective functions [13].

3 Basic Concepts

3.1 Multi-objective Optimization Problem

In the last decade, for classification problem only one objective is considered and with respect to that objective problem is solved. But there are many real-life classification problems for which it is necessary to consider multiple objectives. By nature, classification is multi-objective optimization problem. The main intricacy with multi-objective optimization is that there is no clear definition of optimum solution, so it is difficult to compare one solution with another. Multi-Objective

Optimization Problem (MOOP), is stated as finding the value of solution vector y which is set of n decision variables which must satisfy some constraints such that the *M* objective functions are optimized. Multi-objective optimization problem in mathematical form [3],

$$\begin{array}{lll} \text{Maximize/Minimize} & f_m(y), & m = 1, 2, \ldots, M; \\ \text{Subject to} & g_j(y) \geq 0 & j = 1, 2, \ldots, J; \\ & h_k(y) = 0 & k = 1, 2, \ldots, K; \\ & y_L^i \leq y_i \leq y_U^i & i = 1, 2, \ldots, n \end{array}$$

where,

$g_j(y) \geq 0$ and $h_k(y) = 0$ are inequality constraints and equality constraints respectively. y is vector of n decision variables: $y = (y_1, y_2, y_3, \ldots, y_n)^T$.

3.2 Multi-objective Evolutionary Algorithms

In literature, evolutionary algorithms were used for solving multi-objective problems. There are different multi-objective evolutionary algorithms available for solving multi-objective problems [3, 14]. Different multi-objective evolutionary algorithms are Non-dominated Sorting Genetic Algorithm [15], Strength Pareto Evolutionary Algorithm [16], Strength Pareto Evolutionary Algorithm II [17], Pareto Archived Evolution Strategy [18], Pareto Envelope-based Selection Algorithm [19], Pareto Envelope-based Selection Algorithm II [20] and NSGA II [8].

4 Proposed Framework for Hybrid Non-dominated Sorting Genetic Algorithm II Using Backpropagation

In this paper, hybrid NSGA II is used to evolve neural network by simultaneously optimizing Accuracy and Mean Squared Error using local search. For classification problems we can consider different objectives like accuracy, specificity, sensitivity, precision, mean squared error and recall. In this paper, Accuracy and Mean Squared Error are considered as objectives for optimization.

Maximize Accuracy

$$\text{Accuracy}(A)[2] = \frac{(\text{TP} + \text{TN})}{(T + N)} \tag{1}$$

Minimize Mean Squared Error

$$\text{Mean Squared Error}(E)[2] = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y}_i)^2 \tag{2}$$

Evolutionary algorithms are computationally expensive, i.e. it requires longer search time and evolutionary approach becomes slow [8]. For proposed work hybrid approach is used to speed up the slow convergence by using local search technique with evolutionary algorithm. For proposed work, we have used back-propagation as local search algorithm. Local search algorithm finds the best solutions in small search space. Detailed explanation for Hybrid NSGA II framework is given below:

Initially, random population of size N is generated for first generation. Population of size N is evaluated based on Accuracy and Mean Squared Error and find the fitness of each individual in random population. Assign rank to each individual according to how many solutions it dominates. After assigning ranks to current population, assign crowding distance to each individual which represent crowdedness of the individual. Binary Tournament selection method is used to select parent individuals for reproduction of offsprings on the basis of non-domination rank and crowding distance. Crossover and mutation [9] is used to produce offspring population which is used as population for next generation. Evaluate the offspring population based on Accuracy and Mean Squared Error and find fitness for each individual in offspring population. Apply local search, i.e. backpropagation on combined population, i.e. Parent and offspring population. Do non-dominated sorting and estimate crowding distance of optimized population which is an output of backpropagation. Select top N individuals for next generation. If stopping criteria for algorithm is met then stop next generation and return non-dominated solutions. When maximum generations are reached non-dominated solutions are returned as final solutions (Fig. 1).

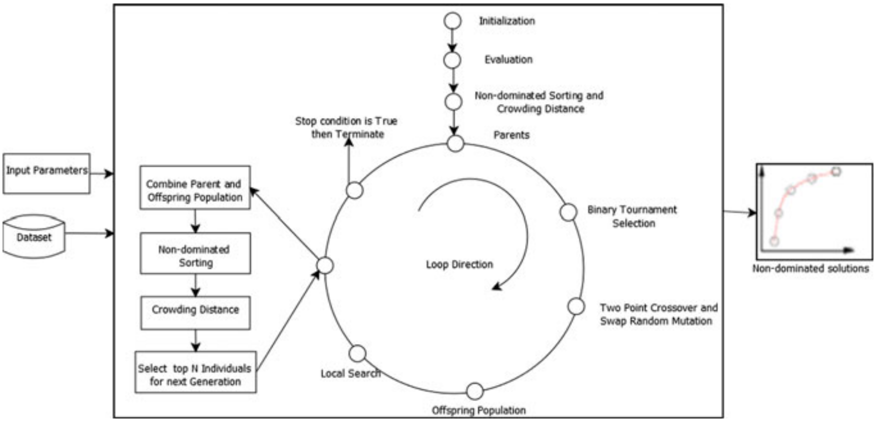

Initialization
Evaluation
Non-dominated Sorting and Crowding Distance
Parents
Stop condition is True then Terminate
Input Parameters
Dataset
Combine Parent and Offspring Population
Non-dominated Sorting
Crowding Distance
Select top N Individuals for next Generation
Binary Tournament Selection
Loop Direction
Two Point Crossover and Swap Random Mutation
Local Search
Offspring Population
Non-dominated solutions

Fig. 1 Proposed framework for Hybrid non-dominated sorting genetic algorithm II [11]

5 Dataset Description and Results

For our work, we considered 11 datasets from UCI (University of California, Irvine) repository [21]. The following datasets are used for proposed work: Breast cancer dataset with 10 attributes and 699 instances, Contact lenses dataset with 5 attributes and 24 instances, CPU dataset with 7 attributes and 209 instances, Diabetes Diagnosis dataset with 9 attributes and 768 instances, Glass dataset with 10 attributes and 214 instances, Heart dataset with 14 attributes and 303 instances, Iris dataset with 5 attributes and 50 instances, Liver disorder dataset with 7 attributes and 345 instances, New thyroid dataset with 6 attributes and 215 instances, Weather dataset with 5 attributes and 14 instances and Wine dataset with 14 attributes and 178 instances.

5.1 Results

In [22] proposed multi-objective approach to evolve neural network using micro-hybrid genetic algorithm. Accuracy of HMON for four datasets is given in Table 1. As the outcome of hybrid NSGA II is set of Pareto fronts, i.e. non-dominated solutions. This Pareto front indicates set of neural network models. The graph in Fig. 2 shows the accuracy of Hybrid NSGA II is better than HMON. For HMON

Table 1 Accuracy results

Datasets	HMON [22] (%)	NSGA II (%)	Hybrid NSGA II (%)
Diabetes diagnosis	75.36	79.43	80.73
Breast cancer	96.82	96	97.99
Heart	81.06	91.6	93.2
Iris	91.03	98.2	98.4

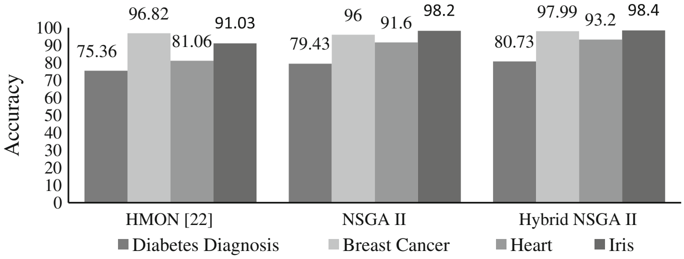


Fig. 2 Accuracy of Hybrid NSGA II compared with HMON [22]

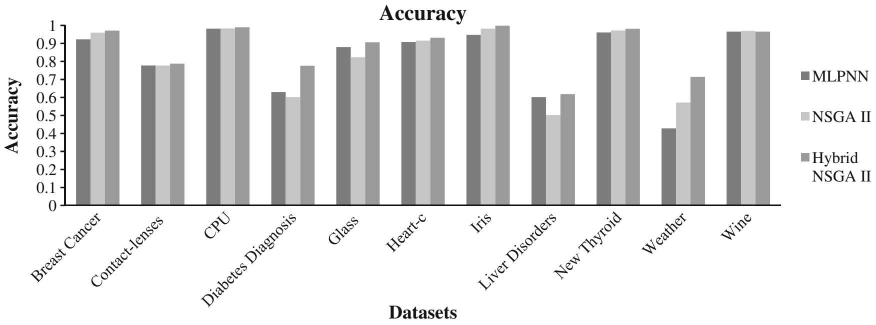


Fig. 3 Accuracy of Hybrid NSGA II compared with NSGA II and MLPNN

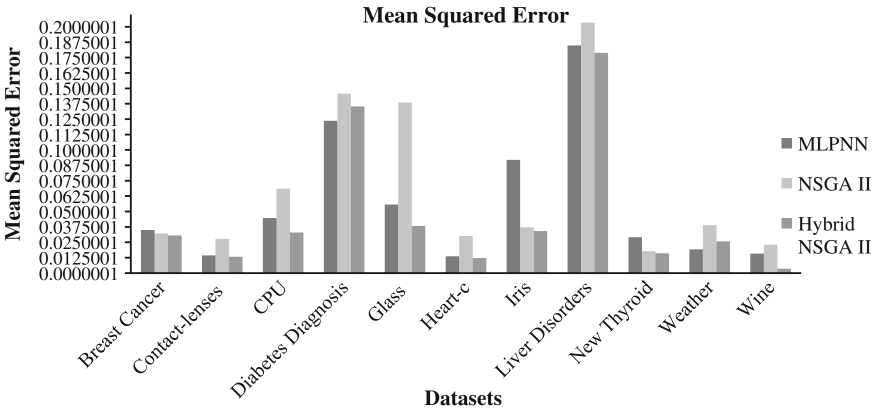


Fig. 4 Mean squared error of Hybrid NSGA II compared with NSGA II and MLPNN

implementation author has considered four datasets so we used four datasets for comparison on the basis of Accuracy.

Figure 3 shows Accuracy comparison of Hybrid NSGA II, NSGA II and Multilayer perception neural network (MLPNN) for different datasets. Graph in Fig. 3 shows that accuracy of hybrid NSGA II for different datasets is better than NSGA II and MLPNN.

In Fig. 4 we compared the results of Hybrid NSGA II with NSGA and MLPNN on the basis of Mean Squared Error (MSE). From graph we can say that for all datasets MSE for Hybrid NSGA II is smaller than MLPNN and NSGA II.

6 Conclusion and Future Scope

In this paper, we have presented hybrid EMO framework for classification problem. This approach is used to optimize Accuracy and Mean squared error simultaneously of classification problem. This hybrid framework brings out faster convergence to Pareto optimal solutions using local search technique. Diversity in solutions is maintained by estimating crowding distance of each individual. This approach preserves best solutions using elitism as best solutions may get lost due to genetic operators. Future research can be to use multi-objective optimization approach for unbalanced dataset.

References

1. U. Maulik, A. Mukhopadhyay and S. Bandyopadhyay: Multiobjective Genetic Algorithms for Clustering—Applications in Data Mining and Bioinformatics. Springer, Berlin, Germany (2011).
2. Han and Kamber: Data Mining: Concepts and Techniques. San Francisco, CA, USA: Morgan Kaufmann (2006).
3. Kalyanmoy Deb: Multi-Objective Optimization Using Evolutionary Algorithms. ISBN 8126528044, 9788126528042, Wiley India (2010).
4. M. Butler and A. Daniyal: Multi-objective optimization with an evolutionary artificial neural network for financial forecasting. In: GECCO'09, 978-1-60558-325-9/09/07, pp. 1451-1457, ACM (2009).
5. K.O. Stanley and R. Mikkulainen: Evolving neural networks through augmenting topologies. Evolutionary Computation, 10(2) (2002).
6. H. Abbass: An evolutionary ANN approach for breast cancer diagnosis. In: Journal Artificial Intelligence in Medicine, vol. 25, Issue 3, pp. 265-281, Elsevier Science (2002).
7. Abbaas HA, Sarker R and Newton C: A Pareto-differential evolution approach for multi-objective optimization problems. In: IEEE congress on evolutionary computation, vol 2, pp. 971-978 (2001).
8. K. Deb, S. Agrawal, T. Meyarivan and A. Pratap: A fast and elitist multiobjective genetic algorithm: NSGA-II. In: IEEE Trans. Evol. Comput., vol. 6, no. 2, pp. 182–197, IEEE (2002).
9. A. Ibrahim, S. Shamsuddin, N. Ahmad and S. Qasem: Multi-objective genetic algorithm for training TTBPN. In: Proc. 4th Int. Conf. on Computing and Informatics (ICOCI 2013), pp. 32–38 (2013).
10. J. Ang, C. Goh, E. Teoh and A. Mamun: Multi-objective evolutionary recurrent neural networks for System Identification. pp. 1586–1592, IEEE (2007).
11. R. Furtuna, S. Curteanu and Leon: An elitist NSGA II algorithm enhanced with a neural network applied to the multi-objective optimization of a polysiloxane synthesis process. In: Journal Engineering Applications of Artificial Intelligence, vol. 24, Issue 5, pp. 772–785, Elsevier (2011).
12. M. Cruz-Ramirez, C. Hervas-Martinez, J. Fernandez, J. Briceno and M. Mata: Predicting patient survival after liver transplantation using evolutionary multi-objective artificial neural networks. In Journal Artificial Intelligence in Medicine, vol. 58, Issue 1, pp. 37–49, Elsevier (2013).
13. D. Gossard, B. Lartigue and F. Thellier: MOO of a building envelope for thermal performance using GAs and ANN. In: Journal of Energy and Buildings 67, pp. 253–260, Elsevier (2013).

14. C. Coello, D. Veldhuizen and G. Lamont: Evolutionary Algorithms for Solving Multi-Objective Problems. Gen. and Evol.Computation, 2nd ed., ISBN 0387367977, 9780387367972, Springer, Berlin/Heidelberg, Germany (2007).
15. K. Deb and N. Srinivas: Multiobjective optimization using nondominated sorting in genetic algorithms. In: Journal Evol. Comput., vol. 2, pp. 221–248 (1994).
16. L. Thiele and E. Zitzler: Multiobjective evolutionary algorithms: The strength Pareto approach. In: IEEE Trans. Evol. Comput., vol. 3, pp. 257–271 (1999).
17. E. Zitzler, L. Thiele and M. Laumanns: SPEA2: Improving the strength Pareto evolutionary algorithm. In: Proc. EUROGEN Int. Conf. on Evol. methods for Design, Opti. And control with appls. to industrial problems, pp. 95–100 (2001).
18. J. Knowles and D. Corne: The Pareto archived evolution strategy: A new baseline algorithm for Pareto MOO. In: Proc. IEEE Cong. Evol. Comput., pp. 98–105, IEEE (1999).
19. D. Corne, M. Oates and J. Knowles: The Pareto envelope based selection algorithm for multiobjective optimization. In: Proc. Conf. PPSN-VI, pp. 839–848 (2000).
20. D. Corne, J. Knowles, N. Jerram, and M. Oates: PESA-II: Region-based selection in evolutionary multiobjective optimization. In: Proc. GECCO, pp. 283–290 (2001).
21. University of California, Irvine (UCI) Repository of Machine Learning Databases https://archive.ics.uci.edu/ml/machine-learning-databases/.
22. C. Goh, E. Teoh, and K. C. Tan: Hybrid multi-objective evolutionary design for artificial neural networks. In: IEEE Trans. On neural networks, vol. 19, pp. 1531–1548, IEEE (2008).

Effectiveness of Email Address Obfuscation on Internet

Prabaharan Poornachandran, Deepak Raj, Soumajit Pal and Aravind Ashok

Abstract Spammers collect email addresses from internet using automated programs known as bots and send bulk SPAMS to them. Making the email address difficult to recognize for the bots (obfuscate) but easily understandable for human users is one of the effective way to prevent spams. In this paper, we focus on evaluating the effectiveness of different techniques to obfuscate an email address and analyze the frequency at which spam mails arrive for each obfuscation technique. For this we employed multiple web crawlers to harvest both obfuscated and non-obfuscated email addresses. We find that majority of the email addresses are non-obfuscated and only handful are obfuscated. This renders majority of email users fall prey to SPAMS. Based on our findings, we propose a natural language processing (NLP)-based obfuscation technique which we believe to be stronger than the currently used obfuscation techniques. To analyze the frequency of arrival of spam mails in an obfuscated mail, we posted obfuscated email addresses on popular websites (social networking and ecommerce sites) to analyze the number of spams received for each obfuscation technique. We observe that even simple obfuscation techniques prevent spams and obfuscated mails receive less spam mails than the non-obfuscated ones.

Keywords SPAMS · Natural language processing · Internet

P. Poornachandran (✉) · S. Pal · A. Ashok
Amrita Center for Cyber Security,
Amrita University, Amritapuri Campus, Kollam, Kerala, India
e-mail: praba@amrita.edu

S. Pal
e-mail: soumajit@am.amrita.edu

A. Ashok
e-mail: aravindashok@am.amrita.edu

D. Raj
Department of Cyber Security System & Networks, Amrita University,
Amritapuri Campus, Kollam, Kerala, India
e-mail: deepakraj250@gmail.com

H.S. Saini et al. (eds.), *Innovations in Computer Science and Engineering*,
Advances in Intelligent Systems and Computing 413,
DOI 10.1007/978-981-10-0419-3_22

1 Introduction

Any person possessing an email address might have received emails asking him to provide his personal information or follow a link, mostly malicious, or seeking help for a financial transaction as discussed by Smith [1] or advertisement of products or services. A report from Washington post [2] confirms that responding positively to such fraudulent emails usually end up in huge financial loss. Any email messages that are not requested by the user can be classified as spams and the person behind sending spam is called a spammer. Spammers collect email addresses in large quantity and run automated programs to send bulk emails to those addresses. Usually, spammers collect email addresses by crawling popular webpages, purchasing from email vendors or social networking sites, etc. Mostly spams are used for either financial crimes or marketing purposes.

One out of every six email addresses reach spammers [3]. Spam statistics report by Kaspersky Lab [4] reveals that in the third quarter of 2013, total spam constitutes 68.3 % of total email traffic. It is evident from the work by Polakis et al. [5] that spammers can collect huge number of email addresses by employing different advanced methods.

Spam not only causes inconvenience to users but also deteriorates quality of internet and possesses serious threat [6–9]. A foolproof mechanism to avoid spam is by preventing the spammers to reach the users email address. So obfuscate the email address before publishing it on internet. Obfuscating an email address means making it difficult for the bots to recognize that it is an email address, whereas human users can easily recognize it. Project Honeypot [10] explains different techniques to obfuscate email addresses.

This paper makes the following contributions:

- We study the robustness of the current obfuscation mail techniques and how easy to exploit them. For this, we crawl half a million domains on internet and extract email addresses by developing an automated system that is capable of identifying both obfuscated and non-obfuscated email addresses.
- We propose a novel natural language processing (NLP)-based obfuscation technique which we believe is much stronger than any other current techniques employed.
- We examined the rate at which obfuscated email address receives spam compared to non-obfuscated emails. Our experiments shows that spammers do not employ any mechanism to identify obfuscated mails, as no mails were received in any of the obfuscated mails posted in popular websites.

2 Literature Survey

Shue et al. [11] post many non-obfuscated email addresses on different popular websites to study how fast email addresses receive spam. They also conduct a study on different ways of harvesting email addresses. They arrived at three main conclusions. Email addresses are harvested quickly by the spammers, crawlers can be tracked and finally, spammers use multiple harvesting techniques. A single email address exposed on internet results in instant and high volume of spam.

Polakis et al. [5] discuss traditional and advanced methods of harvesting email addresses. Traditional methods include web crawling, crawling archive sites, dictionary attack, etc. Advanced methods include blind harvesting, in which names are collected from social networking sites and are searched in Google to extract email addresses in the result page, and targeted harvesting, in which email addresses are harvested using public information available on social networking sites. The study shows that blind harvesting yields more number of email addresses, whereas targeted harvesting collects accurate email addresses with more personal information.

In [12] Li Zhuang et al. studied operation of bots using spam emails, identifies common characteristics of spams, and estimates the size of bots and its geographical distribution. Furthermore, analysis was done by Prince et al. [13] using the data gathered from project Honeypot (www.projecthoneypot.org). Even though there are different techniques [14–16] to fight spams, none of them perfectly differentiate legitimate emails from spams [17] or cause burden to users.

3 Obfuscation Techniques

Project Honeypot [10] lists different obfuscation techniques applicable to email addresses to evade bots. These are the obfuscation techniques we use in this study.

(a) *Random text insertion*: Inserting random text prevents bots from successfully sending emails where as a human user omits the random text before sending an email. For example: *userDELETETHIS@example.com* will be sent instead of *user@example.com*. Replacing com with zom is a similar technique used. e.g., *user@gmail.zom*.

(b) *Replacing symbols with words*: Symbols of an email address are replaced with corresponding words or other characters or both. For example, *user(at)example(dot)com*, *user()gmail!com,* etc. Guessing such characters or including every possible character in a bot is yet another challenge for spammers.

(c) *Using ASCII code*: Either the entire email address or only the ‘@’ or ‘.’ symbol is replaced with corresponding ASCII code. E.g., *user&#64;example&#46;com*. When a browser executes it, the ASCII codes are converted to corresponding ‘@’ and ‘.’ Symbols.

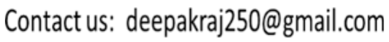
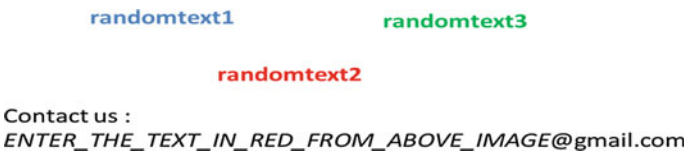

Contact us: deepakraj250@gmail.com

Fig. 1 Email address obfuscation technique by embedding it in an image


randomtext1 randomtext3

randomtext2

Contact us :
ENTER_THE_TEXT_IN_RED_FROM_ABOVE_IMAGE@gmail.com

Fig. 2 NLP based email address obfuscation technique

(d) *JavaScript Obfuscation*: Uses JavaScript code to split the email address and store it in an array. Then each array is concatenated to display the email address. For, e.g., joe.smith@example.com is obfuscated as follows; *<script type='text/javascript'>var a=new Array('com','le.', 'ith','.sm','joe','@ex','amp');document.write("<a[4]+a[3]+a[2]+a[5]+a[6]+a[1] +a[0]+"</a>");</script>*

(e) *E-mail address in an image*: Display an image of the email address instead of text. Reading text from image is complex and time consumingas shown in Fig. 1.

(f) *NLP-based Obfuscation*: Here, we move one step forward than just keeping the email addresses in the image. The idea is to keep multiple texts in the image and provide a linguistic question/hint through which only humans would be able to identify the exact text to be taken from the image. One of the examples of NLP-based obfuscation is shown in Fig. 2.

4 Methodology

This paper was motivated by two research questions. "How much effective each obfuscation technique is?" and "How fast obfuscated email addresses are being identified by bots?" To answer first question, both obfuscated and non-obfuscated email addresses were collected from internet. Then they were grouped according to the obfuscation technique to analyze the number of email address received in each obfuscation technique. To answer second research question, several email addresses were created, obfuscated using different techniques, and posted them on popular webpages to analyze the number of spams received for a period of 4 months.

4.1 Harvest Email Address from Internet

The main objective of this step was to study the effectiveness of obfuscation technique used to make email address difficult to understand by bots. We crawled

Alexa Internet [18] to extract top 500,000 domains. Then each domain was crawled to obtain all the obfuscated and non-obfuscated email addresses in it. We followed the naive approach of scanning each line and matching it against a set of regular expressions pertaining to different type of obfuscation techniques [10]. A non-obfuscated email address (user@gmail.com) could be identified by matching the text against the regular expression *[A-Z0-9._%+−]+@[A-Z0-9.-]+[A-Z]{2,4}.*

The regular expression searches for any text followed by '@' character. Then continues to search for remaining text in that word followed by a '.' character and finally a text of length 2–4. Non-obfuscated email addresses were also harvested in this study to analyze the impact of obfuscation.

Inserting random texts is another obfuscation technique focused in this project. Legitimate experienced users may easily understand how to alter the obfuscated email address but for novice users it may appear bit complex. So it is a good idea to display a note on how to alter the given email address at the end. This type of obfuscated email addresses could also be harvested using the above-mentioned regular expression. The main advantage to legitimate users is that the bots treat it as a normal email address, and hence may not take extra effort to remove the inserted random text. So, upon sending spams to such addresses, messages will be bounced by the email server, and hence will not be received by the user (Fig. 3).

Replacing symbols with words is a simple approach to obfuscate. The above-mentioned regular expression should be tweaked a little bit to handle such obfuscation. Instead of '@' and '.' characters, regular expression should match at and dot words. Incorporating the combination of all possible symbols into bot is a challenge. An alternative approach for the spammers is that they identify the obfuscation on a particular website first and then modify their program accordingly. But this limits the degree of automation.

The regular expression should be altered in such a way that instead of '@' and '.' characters it should match the ASCII codes, *&#64;* and *&#46;* to harvest ASCII

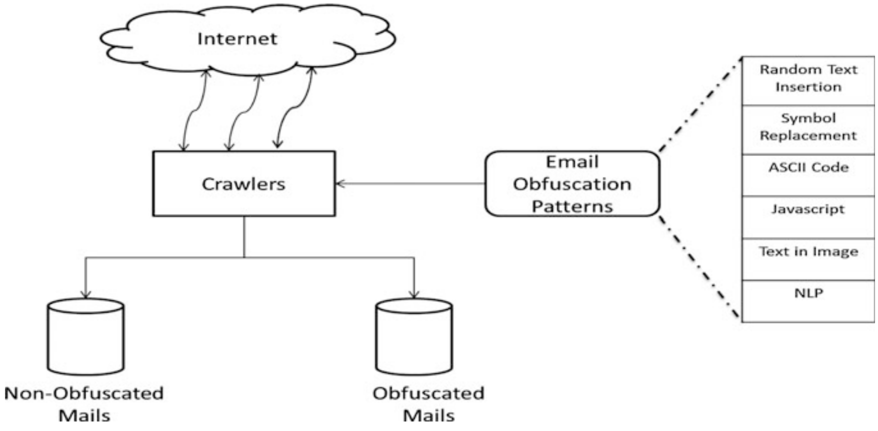


Fig. 3 Harvesting email addresses from the internet

encoded addresses. User can also convert the entire address to corresponding ASCII characters to ensure more complexity.

JavaScript obfuscation is another complex technique used to obfuscate email addresses. We searched for keywords like '*com,*' '@' or everything pertaining to an email address. If any match is found, then the entire code is saved. As we got only a handful of such code we analyzed the code manually to identify any email address, but there were no valid one. At a larger scale it is difficult to automate this process. Users may combine other obfuscation techniques like replacing '@' with '*at,*' etc. The main disadvantage of this technique is that the browser should be JavaScript compatible to run this code. So an alternate obfuscation technique should also be given along with this script to handle browser with no JavaScript support.

These were the obfuscation techniques analyzed in this study. Other techniques like embedding address on an image or the NLP-based obfuscation is more complex and spammers are unlikely to invest huge amount of time in today's scenario. A system was dedicated to run the process continuously till it scans 500,000 domains. In this approach, we grouped 500,000 web domains into two where the first group contains web domains pertaining to social networking sites, blogs, mailing lists, etc. In the first group, it is likely that probability of finding email addresses in every page is high. So every page in each domain of this group is scanned. The second group contains list of websites that host information about particular company, services, etc. In the second group, we scan only those pages where the contact information is displayed. Usually, these pages will be under the name of contact, contact info, about us, etc. We assume that the same obfuscation technique will be maintained throughout the domain. A naive approach like this itself harvested two and a half millions of non-obfuscated email addresses in a short span of time. So the magnitude of email addresses harvested would be very high considering the advanced technologies employed by the modern day spammers. After the completion of data collection, analysis of the collected data was carried out. More details of the analysis are shown in the subsequent section.

4.2 Publish Email Addresses on Internet

Main objective of this step was to find out how fast each obfuscated email addresses reach spammers. The general approach was to publish obfuscated email addresses on internet and record the number of spams received in each email address. Two questions to be answered in this section are: "How early each obfuscated email addresses received spam?" and "How many spams were received by them?" (Fig. 4).

A total of 220 non-obfuscated email addresses were created using web email clients like Gmail, Hotmail, Yahoo, Rediffmail, Yandix, inbox, etc. The 320 email addresses were categorized into six sets as shown in Table 1. 10 non-obfuscated email addresses, one from each email client, were intentionally left unused to check whether any of the email clients send spams or sell email addresses to third parties who send spams. Another 10 non-obfuscated email addresses were used to create account in popular

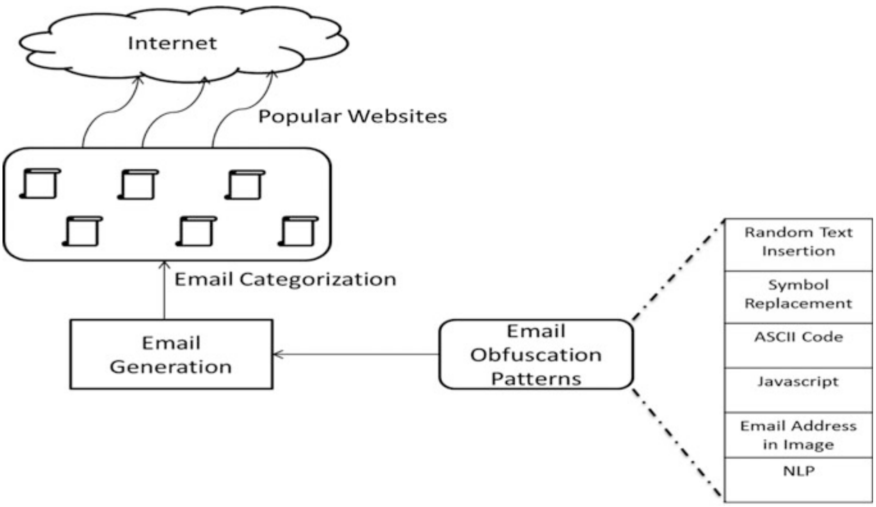


Fig. 4 Publishing email address on internet

Table 1 Overview of the email address sets used

Set no.	Obfuscation type	Example	No. of email address
Set 1	None	*user@gmail.com*	70
Set 2	Inserting random text	*userREMOVE@gmail.com*	50
Set 3	Replace *com* with *Zom*	*user@gmail.zom*	50
Set 4	Replace “@”, “.” with words	*user(at)gmail(dot)com*	50
Set 5	Email inside an image	Figure 1	50
Set 6	NLP-based obfuscation	Figure 2	50

ecommerce applications like Flipkart, snapdeal, eBay, etc. Registering in such sites using obfuscated email addresses like *userDELETETHIS@gmail.com*, where original email address is *user@gmail.com*, was possible. But many legitimate email communications like account verification, etc. were done by automated programs that may not understand the original address. So such legitimate messages will be bounced back and will not reach users. Another 50 non-obfuscated email addresses were used to publish as comments on blogs like Washington post, ecommerce sites, video blogs like YouTube, etc. The exact URL of each comment locations was recorded. Email addresses in other sets too were published as comments in same places as that of set 1. Then email addresses were regularly checked for any spams for a period of 4 months. More detailed analysis of the analysis is shown in the next section.

5 Overview of Data Collection

In this section we analyze the collected data. This section is divided into two parts. Section A is provided with the analysis of harvested obfuscated email addresses and Section B is provided with the analysis of spams received on obfuscated email addresses.

5.1 *Overview of Email Addresses Harvested from Internet*

A total of 2,500,000 email addresses were received on crawling 500,000 domains on internet. It came as a surprise that 2,475,000 of them were non-obfuscated that constitutes 99 % of them. The only obfuscation that could be found was replacing '@' and '.' with other characters, which were 32,337 of them. From the enormous amount of email addresses collected, no other types of obfuscation were found. A single system dedicated for the process using such a naive approach itself could identify around two and a half million email addresses in a short span of time in which only handful were obfuscated. Spammers who employ advanced methods to crawl could easily get millions of email addresses in much lesser time frame. It is evident from the study that majority of the users did not obfuscate their email addresses before publishing it on internet. This makes work of spammers easier as they do not have to make complex programs to harvest large number of email addresses (Fig. 5).

5.2 *Overview of Email Addresses Published on Internet*

In a span of 4 months, a total of 422 spams were received. Our observations show that all of them were on non-obfuscated email addresses. Not a single spam was

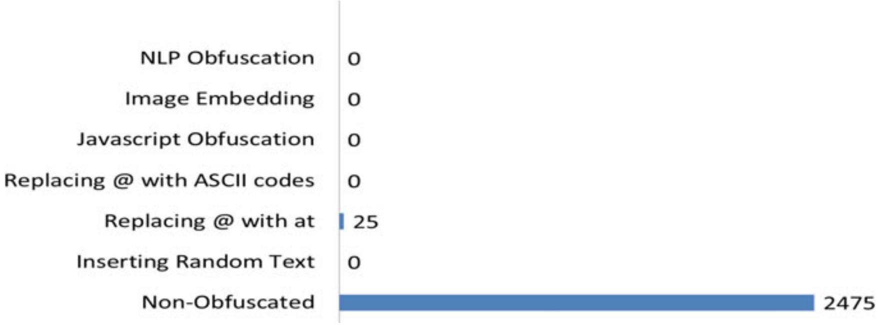


Fig. 5 Email addresses harvested from the internet based on their obfuscation technique

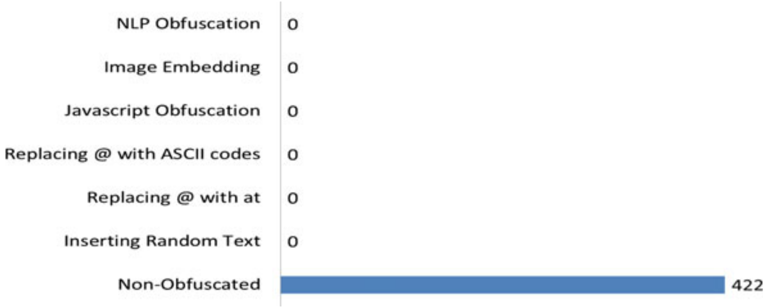


Fig. 6 Spam mails received by both obfuscated and non-obfuscated email addresses

received on obfuscated ones. Out of 422 spams received on non-obfuscated email addresses, first spam received within 4 h after publishing on internet. Majority of the spams were on those email addresses that were published as comments on popular websites. Spams were received in all email addresses within first 24 h on such non-obfuscated email addresses. Spams were received at a rate of 3 per day. Non-obfuscated email addresses published on blogs got more number of spams (Fig. 6).

Even though the study done in the previous section of this paper showed that it is easy to harvest obfuscated email addresses, no spams were received on any of them. To convert an obfuscated email address to its original form is not only more complex program but also more time has to be invested by spammers. Since huge numbers of non-obfuscated email addresses are easily available, investing more time and complex programs seems unnecessary for them and that might have made them to ignore even simplest of the obfuscation technique.

As we have pointed out, all spams were received in non-obfuscated email addresses. None of the addresses used to create accounts in different applications like Flipkart, snapdeal, etc. received any spam. Also none of the addresses that were intentionally left unused receive any spam except one rediffmail account. 12 spams were received on a rediffmail email address that was not published anywhere on the internet and the first one was received within 48 h after creating that address. This leads us to the conclusion that either spammers can successfully harvest addresses from them or companies sell user's personal information like email address to spammers.

The email addresses used were freshly created only for this study. Fifty non-obfuscated ones were published as comments on various locations on internet. Even such a newly created email address, each posted in only one location of internet, received spams. This shows the intensity of spammers at which they send spams. A brief survey among long time email users revealed that they were receiving an average of 20 spams per day. So users may apply at least a simple obfuscation technique to prevent spam for the time being.

The protection offered by such techniques greatly depend on how easily automated programs can extract original form of email address from obfuscated ones.

Replacing '@' with 'at' or its corresponding ASCII code is relatively simple to convert to its original form. But consider the scenario of inserting a random text into an address. For, e.g., bob***@gmail.com and ask the viewer to remove *** from it to get the original form, i.e., bob@gmail.com, keep the automated programs guessing on what would be the random text that is not part of the original address since there need not be any specific pattern for such a text.

6 Conclusion

To study the effectiveness of email obfuscation techniques and impact of spam on obfuscated email addresses were the main focus of this paper. We also propose NLP-based email address obfuscation technique which we feel is stronger than any other obfuscation technique employed till date. Study was done in two independent modules, 500,000 domains were crawled, in the first module, to harvest obfuscated and non-obfuscated email addresses. In the second module, many non-obfuscated and obfuscated email addresses were published on internet to study how fast those email addresses were harvested by spammers. Data collected from first module showed that out of 2,500,000 email addresses harvested from the first module, 2,475,000 of them were non-obfuscated and 32, were obfuscated by replacing '@' with 'at'. No other type of obfuscation technique were found in 500,000 domains crawled as part of this study. In the second module, no spams were received on any of the obfuscated email addresses, whereas a number of spams were received on non-obfuscated ones.

Our experiments show that obfuscated email addresses that can be identified by simple regular expression are relatively easier to harvest. From the second module it is clear that obfuscated email addresses prevent spams very effectively. But this protection greatly depends upon on how common obfuscation is in the coming days. Automating the conversion of obfuscated email address to its original form is another challenge spammers may face for, e.g., guessing the random text inserted or the symbol used to replace @ in an email address. It is highly likely that the day spammers do not get huge number of email addresses, they start to harvest obfuscated email addresses. Even then reading email addresses from images will remain as a challenge as scanning entire images on internet is cumbersome.

References

1. Andrew Smith "Nigerian scam e-mails and the charms of capital" Cultural Studies #160; 23, 1, 27–47.
2. Washington post blog: Fell for That Nigerian Scam? See Him. http://www.washingtonpost.com/wp-dyn/content/article/2006/04/19/AR2006041901162.html.
3. Blog: Email Tips to Avoid Spam Filters and 9 Top Email Marketing Tools http://blog.woorank.com/2013/01/email-tips-to-avoid-spam-filters-and-9-top-email-marketing-tools/.

4. Kaspersky Blog: Spam statistics Report Q3-2013. http://usa.kaspersky.com/internet-security-center/threats/spam-statistics-report-q3-2013#.U3Igpd-bAjw.
5. Iasonas Polakis, Georgios Kontaxis, Spiros Antonatos "Using Social networks to Harvest Email Addresses", in WPES '10: Proceedings of the 9th annual ACM workshop on Privacy in the electronic society.
6. Allister Cournane, Ray Hunt "An analysis of the tools used for the generation and prevention of spam", Computers & Security, vol. 23, pp. 154–166, 2004.
7. Vidyasagar Potdar, Nazanin Firoozeh, Farida Ridzuan, Yan Like, Debajyoti Mukhopadhyay, Dhiren Tejani "The Changing Nature of Spam 2.0".
8. Deborah Fallows "Spam: How it is hurting email and degrading life on the Internet", *Pew Internet & American Life Project report. Retrieved August 10, 2007 from* http://www.pewinternet.org/PPF/r/102/report_display.asp.
9. Adam Massof "Spam - Oy, What a Nuisance".
10. Project honey pot: Help stop spammers before they ever get your address! http://www.projecthoneypot.org/.
11. Craig A. Shue, Minaxi Gupta, Chin Hua Kong, John T. Lubia, Asim S. Yuksel "Spamology: A Study of Spam Origins", in the 6th Conference on Email and Anti-Spam (CEAS) (2009).
12. Li Zhuang, John Dunagan, Daniel R. Simon, Helen J. Wang, Ivan Osipkov, Geoff Hulten, J. D. Tygar "Characterizing bots from email spam records", In *Proceedings of the First USENIX Workshop on Large-Scale Exploits and Emergent Threats (LEET'08)*, 2008.
13. Matthew B. Prince, Lee Holloway, Eric Langheinrich, Benjamin M. Dahl, Arthur M. Keller, "Understanding How Spammers Steal Your E-Mail Address: An Analysis of the First Six Months of Data from Project Honey Pot", in Conference on Email and Anti-Spam (CEAS), 2005
14. John R. Levine "Experiences with Greylisting" Proc. 2nd Conf. Email and Anti-Spam (CEAS 05), 2005; www.ceas.cc.
15. Ion Androutsopoulos, John Koutsias, Konstantinos V. Chandrinos, George Paliouras, Constantine D. Spyropoulos "An Evaluation of Naive Bayesian Anti-Spam Filtering". In Proceedings of the Workshop on Machine Learning in New Information Age, Barcelona, Spain, 2000.
16. Cynthia Dwork, Moni Naor "Pricing via Processing or Combatting Junk Mail". In E.F. Brickell, editor, Advances in Cryptology—Crypto'92, pages 139–147, 1992.
17. Hallam-Baker, P. (2003). "A plan for no spam". Technical report, Verisign.
18. Web Information Service "Alexa Internet", http://www.alexa.com.

A Proof-of-Concept Model for Vehicular Cloud Computing Using OMNeT++ and SUMo

Mohammad Pasha, Mohd Umar Farooq and Khaleel-Ur-Rahman Khan

Abstract Vehicular cloud computing (VCC) adapts from the fact that vehicular nodes can use their on-board computational power, storage, and communication resources to interact with the Cyber-Physical elements. Through this study we identify VCC as an essential stepping stone toward visualizing the Internet of Vehicles (IoV) ecosystem to integrate mobile ad hoc networks, wireless sensor networks, mobile computing, and cloud computing. We provide a classification on the state of the art of VCC by drawing a relationship between the existing domains and IoV through a review of the important works carried recently in the literature. Finally, we present a proof of concept model for vehicular clouds and propose a vehicular resource discovery protocol and evaluate it through simulations using OMNeT++ and SUMo.

Keywords Internet of things · Internet of vehicles · Vehicular cloud computing · Mobile cloud computing · Vehicular ad hoc networks · Wireless sensor networks · Cloud computing · Mobile computing

1 Introduction

Vehicular cloud computing (VCC) can be seen as a combination vehicular networks and cloud computing. The vehicular ad hoc network (VANET) primarily adapts from mobile ad hoc networks (manets) where communication among the nodes is of

M. Pasha (✉) · M.U. Farooq
MJCET, Hyderabad, India
e-mail: muhammed.pasha@gmail.com

M.U. Farooq
e-mail: umarfarooq.mohd@gmail.com

K.-U.-R. Khan
ACE, Hyderabad, India
e-mail: khaleelrkhan@gmail.com

H.S. Saini et al. (eds.), *Innovations in Computer Science and Engineering*,
Advances in Intelligent Systems and Computing 413,
DOI 10.1007/978-981-10-0419-3_23

prior importance and is generally single hop. This component is aimed to assist and improve the intelligent transport systems. The cloud computing perspective of the VCC model adapts from mobile cloud computing where services provided by remote servers are accessed in a way similar to the telecommunication. A hybrid of the above models can be a cloud computing model that oscillates between being a pure vehicular cloud or semi-vehicular cloud which can executes jobs in a similar fashion to a conventional cloud providing "as a service" models. This is not possible in case of resource constrained devices like smartphones or other sensing devices that have constrained battery life and limited communication range. The contribution of this study is to identify VCC as an essential stepping stone toward visualizing the Internet of Vehicles (IoV) eco-system to integrate MANETs, wireless sensor networks, mobile computing, and cloud computing as seen in Fig. 1. Furthermore, we envision VCC as a supplement to Conventional Cloud Computing for on road services.

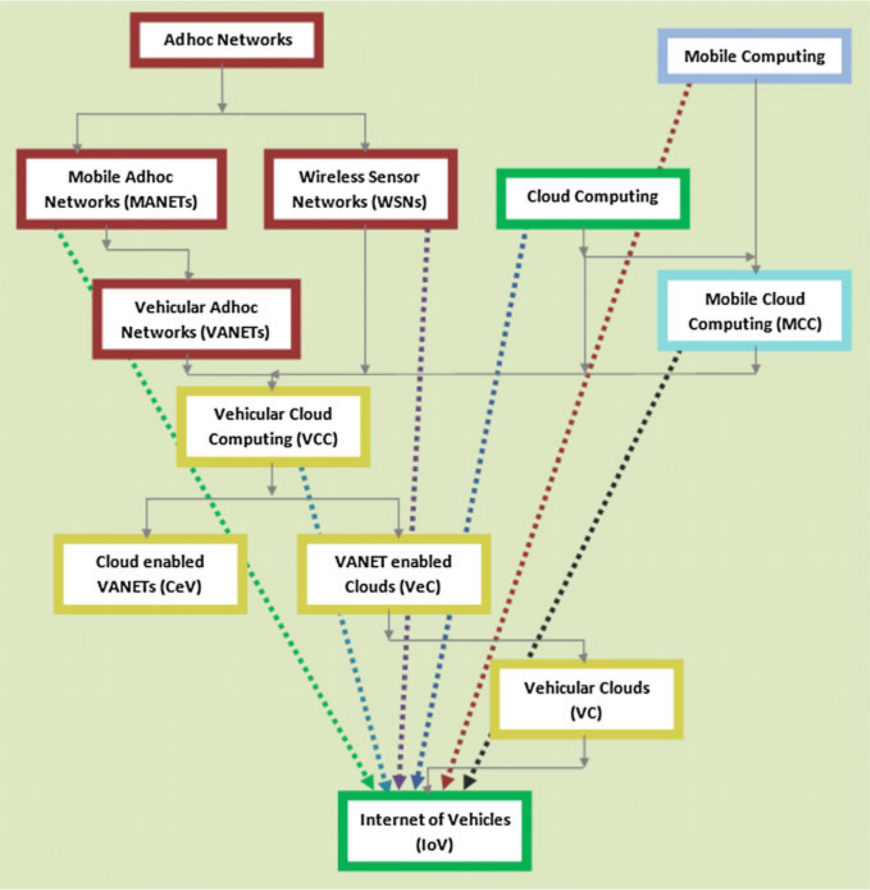

Adhoc Networks
Mobile Computing
Mobile Adhoc Networks (MANETs)
Wireless Sensor Networks (WSNs)
Cloud Computing
Vehicular Adhoc Networks (VANETs)
Mobile Cloud Computing (MCC)
Vehicular Cloud Computing (VCC)
Cloud enabled VANETs (CeV)
VANET enabled Clouds (VeC)
Vehicular Clouds (VC)
Internet of Vehicles (IoV)

Fig. 1 Technologies contributing towards Internet of Vehicles

2 Vehicular Cloud Computing and Internet of Vehicles

The concept of IoV is an integral part of Internet of Things. Cloud computing is seen as to fuel the infrastructure that runs the Internet of Things [1]. VCC is proposed to enable the interaction of vehicular nodes with the environment, i.e., roads, residential buildings, market places, etc., through a collection of sensors that can gather useful information which can be utilized to build an efficient part of the ecosystem that constitutes the Internet of Things, i.e., The IoV. The sort of information gathering will not be possible through other means. This natural extension of the vehicular networks to use the cloud-based service offerings can help in exhaustive information gathering, analytics, and computations to evaluate the socio-environmental effects of the Internet of Things.

3 The Emergence of Vehicular Cloud Computing

The modern day vehicular networks need to evolve into tomorrow's IoV through a substantial transformation in three different directions namely client-connection-cloud [2]. The client component represents the resources inside a vehicle that can locally store, compute, and disseminate useful information to the surroundings. The connection models the communication technologies to be used as a medium for information interchange primarily between the clients. The cloud system provides the infrastructure needed for performing the heavy duty operations that cannot be fulfilled by the constrained resources available inside the vehicle. It also connects the vehicular ecosystem to the components of the Internet of things paradigm.

4 Literature on Modeling Vehicular Cloud Computing

4.1 The Autonomous Vehicular Clouds

Lee et al. [3] proposed the autonomous vehicular clouds as "a group of largely autonomous vehicles whose corporate computing, sensing, communication and physical resources can be coordinated and dynamically allocated to authorized users." These autonomous vehicular clouds are seen to serve users with resources like storage carried by the vehicular nodes through integration with the remote infrastructure-based service offerings paradigm, the cloud computing. The author exemplifies the applications of autonomous vehicular clouds in traffic management, assessment management, and other fields. The proposed work foresees vehicular resources as to a supplement to conventional cloud-based offerings.

4.2 A True Vehicular Cloud

Authors in [4] classified mobile clouds as either be Mobile vehicular cloud, mobile personal clouds, and mission-oriented mobile clouds. The research challenges pointed for mobile clouds include Privacy and security protection, sensor filtering and aggregation and content-based, secure networking. The Mobeyes system [5] an Internet connected mobile vehicle cloud to provide safety monitoring-related services using vehicles fitted with cameras was selected to present a proof of concept for VCC. The information carried by the nodes of this system can be traced using protocols designed specifically for vehicular network. Using the example of the traffic management assistance through vehicular clouds the author reveals how timely updated information can be exchanged between mobile vehicular cloud and a central navigation center over internet.

4.3 A Comparison of Models for Vehicular Cloud Computing

The important differences between the VCC proposals of Olariu et al. and Gerla et al. is that the former foresees the under-utilized resources in the vehicular network to supplement the conventional cloud-based offerings while the latter envisions the vehicular nodes to form a self-sufficient intelligent vehicular grid. Also Olariu et al. maintains the use of vehicle to infrastructure communication as a prominent feature for VANET enabled cloud whereas Gerla et al. identified inter-vehicular communications to be upmost importance (Table 1).

5 Proposed Framework for Vehicular Cloud Computing

To model the vehicular cloud we utilized the OMNET++ and the SUMO which offer support to model to two components namely cloud computing and vehicular communications. Extensive literature about these tools and their features can be found in

Table 1 Comparing VANET enabled clouds and cloud enabled VANETs

Feature	VANET enabled cloud	Vehicular cloud
Proposed by	Olariu et al.	Gerla et al.
Idea	Conventional clouds using in-vehicular resources	Intelligent cloud using in in-vehicular resources
Prominent communication technology	Vehicle to infrastructure	Vehicle to vehicle
Layer of cloud computing supplemented	Infrastructure and data center layer	Service, infrastructure and data center layer

[6, 7], respectively. Cloud simulators have already been successfully implemented using OMNeT++. SUMo supports vehicular mobility models which are used for vehicular simulations by a large audience. We used Veins Framework [8] which integrates both these tools to present a proof-of-concept model for VCC. The simulation model consisted of a single lane with vehicles running at varying speeds under urban conditions. In the proposed scenario, vehicles query on-board resources of other vehicles to perform computation offloading. The performance of the protocol for three different strategies namely resource discovery with duplicated messages, resource discovery without duplicated messages and resource discovery based on multistage graph was considered. The parameters considered were the number of entries in the senders list, number of packets forwarded successfully, broadcast-packets received successfully.

It was observed that multistage graph-based strategy out-performed the other techniques by reducing the number of message exchanges during the resource discovery phase (Figs. 2, 3 and 4).

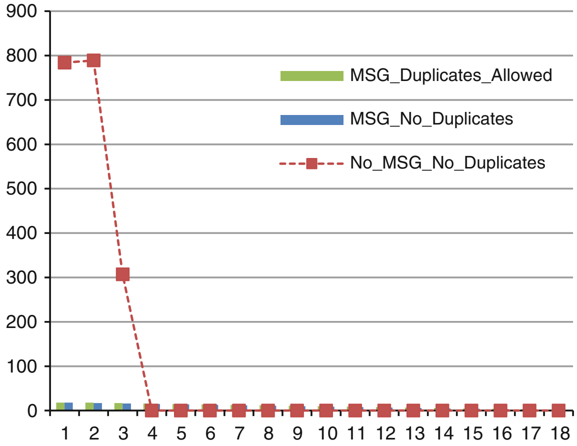


Fig. 2 No. of entries in the senders list

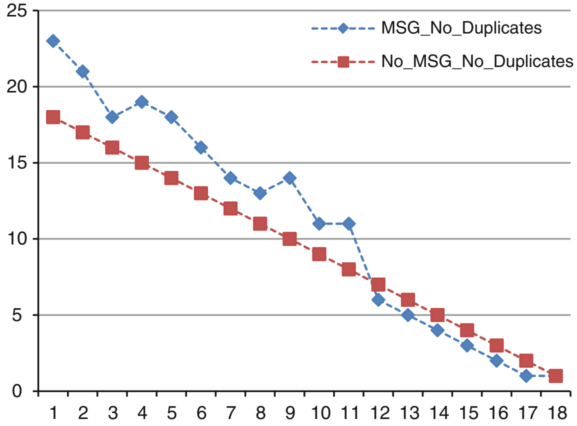
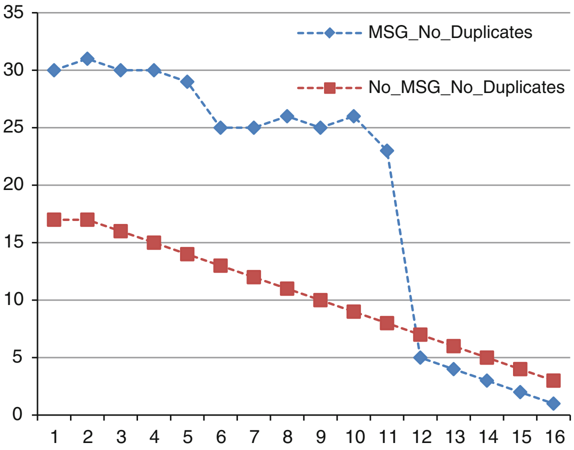


Fig. 3 No. of packets forwarded successfully

Fig. 4 Broadcast packets received successfully



6 Conclusions

We have successfully identified the tools to model and simulate VCC through two well-established tools namely OMNet++ and SUMO. We simulated a vehicular cloud model to discover vehicular resources using a multistage graph-based protocol.

References

1. Gubbi, Jayavardhana, Rajkumar Buyya, Slaven Marusic, and Marimuthu Palaniswami. “Internet of Things (IoT): A vision, architectural elements, and future directions.” Future Generation Computer Systems 29, no. 7 (2013): 1645–1660.
2. Ovidiu Vermesan, Peter Friess, Internet of Things-Converging Technologies for Smart Environment and Integrated Ecosystem, 1st ed., River Publishers, 2013, pp. 08–09.
3. Lee, Euisin, Eun-Kyu Lee, Mario Gerla, and UCLA Soon Y. Oh. “Vehicular Cloud Networking: Architecture and Design Principles.” *IEEE Communications Magazine* (2014): 149.
4. Eltoweissy, Mohamed, Stephan Olariu, and Mohamed Younis. “Towards autonomous vehicular clouds.” In Ad hoc networks, pp. 1–16. Springer Berlin Heidelberg, 2010.
5. Lee, Uichin, Biao Zhou, Mario Gerla, Eugenio Magistretti, Paolo Bellavista, and Antonio Corradi. “Mobeyes: smart mobs for urban monitoring with a vehicular sensor network.” *Wireless Communications, IEEE* 13, no. 5 (2006): 52–57.
6. www.omnetpp.org.
7. www.sumo.dlr.de/wiki/Main_Page
8. Sommer, Christoph, Reinhard German, and Falko Dressler. “Bidirectionally coupled network and road traffic simulation for improved IVC analysis.” *Mobile Computing, IEEE Transactions* 10.1 (2011): 3–15.

Password Reuse Behavior: How Massive Online Data Breaches Impacts Personal Data in Web

Prabaharan Poornachandran, M. Nithun, Soumajit Pal, Aravind Ashok and Aravind Ajayan

Abstract Web 2.0 has given a new dimension to Internet bringing in the "social web" where personal data of a user resides in a public space. Personal Knowledge Management (PKM) by websites like Facebook, LinkedIn, and Twitter, etc. has given rise to need of a proper security. All these websites and other online accounts manage authentication of the users with simple text-based passwords. Password reuse behavior can compromise connected user accounts if any of the company's data is breached. The main idea of this paper is to demonstrate that the password reuse behavior makes one's account vulnerable and these accounts are prone to attack/hack. In this study, we collected usernames and passwords dumps of 15 different websites from public forums like pastebin.com. We used 62,000 and 3000 login credentials from Twitter and Thai4promotion websites, respectively for our research. Our experiments revealed an extensive password reuse behavior across sites like Twitter, Facebook, Gmail, etc. About 35 % of accounts we experimented were vulnerable to this phenomenon. A survey was conducted targeting online users which showed us that, around 59 % out of 79 % regular internet users still reuse passwords for multiple accounts.

Keywords PKM · Password · Web 2.0

P. Poornachandran (✉) · S. Pal · A. Ashok
Amrita Center for Cyber Security, Amrita University, Amritapuri Campus,
Kollam, Kerala, India
e-mail: praba@amrita.edu

S. Pal
e-mail: soumajit@am.amrita.edu

A. Ashok
e-mail: aravindashok@am.amrita.edu

M. Nithun · A. Ajayan
Department of Cyber Security Systems & Networks, Amrita University,
Amritapuri Campus, Kollam, Kerala, India
e-mail: nithunm1@gmail.com

A. Ajayan
e-mail: altoarun@gmail.com

H.S. Saini et al. (eds.), *Innovations in Computer Science and Engineering*,
Advances in Intelligent Systems and Computing 413,
DOI 10.1007/978-981-10-0419-3_24

1 Introduction

Today almost all business runs on the internet. There are different types of accounts based on the business and the business providers. We have email accounts like Gmail, Yahoo mail, etc., we have accounts for shopping/commerce sites like Flipkart, Ebay, accounts for various cloud services like Amazon AWS, Dropbox etc., accounts for content management sites like Wordpress, Blogger and what not. Personal data is pushed to the public domains like Facebook, Twitter, etc. by the users but the common users lack the knowledge of safety of their own data. On an average an average user has more than 25 online accounts on various websites according to Microsoft Research [1, 2]. Most of these websites uses text-based password for the user authentication. The user has to manually enter the username and password for authentication. This method of user authentication has several flaws like, the user has to memorize strong alphanumeric password, the user should be careful of phishing attacks, shoulder sniffing, key loggers, denial of service attacks, etc. [3, 4].

In recent decades, many websites have fallen prey to massive online data breach. So far more than 1.7 billion user account credentials have been leaked. In April 2014 AOL, American multinational mass media corporations' half a million users' accounts were compromised and in October 2013 a much severe attack hammered Adobe, where 38 million user details were compromised. A study reveals that organizations are attacked on an average of 2 million times a week and many of those attacks result in a data breach [5, 6]. Many users practice the reuse of password across multiple accounts in internet. Password reuse allows online attackers to gain access into dozens of accounts used by the individual by compromising one of his accounts [7–10]. Personal knowledge would be no longer safe if any of the users' account falls prey to an online data breach. "thereisnofate-butwhatwemake" a password with 26 characters has been cracked [11–14], so the password strength is not accounted when the database of a company is compromised. All the compromised database dumps are readily available in the internet.In [15, 16] the authors describe about the password re-usability nature employed by normal users and corresponding steps to avoid it.

Many powerful GPGPU based password recovery tools are available which can do millions of guesses per second and these tools can be used for decrypting even the hashed passwords [17–21]. Many hacker groups, such as Anonymous, Lulzsec, Antisec, Chaos Computer Club, TeslaTeam, UGNazi and others are very keen in distributing the breached data across internet and the dumped data becomes handy to all [22–27]. Password reuse initiates a security vulnerability chain because an adversary who compromises one service can compromise other services authenticated by the same password, thus reducing the gross security to that of the weakest site. This is what is exactly conveyed in the phrase "chain is only as strong as its weakest link." Think about the weakest link when a user uses same password for his/her social networking site accounts and online banking sites [28–30].

This paper makes the following contributions:

- We demonstrate that the password reuse behavior makes one's account vulnerable and the user's other accounts linked with the same password are prone to attack/hack.
- We crawled username and password dumps of multiple sites from public forums and databases for studying the password reuse behavior across multiple sites.
- We categorized all the passwords based on their strength. Our experiments prove that even if a user uses the same/similar strong password for a chain of accounts, his accounts are equally vulnerable if any one of them gets hacked.
- We conducted a survey targeting student, staffs, and professionals among multiple universities to understand the password management among users.

The rest of the paper is organized as follows. The data collection and crawling part are mentioned in Sect. 2. We provide a detailed explanation of our analysis and implementation details in Sect. 3. In Sect. 4 we elaborate the experimental results obtained. The survey details and results are explained in Sect. 5. We conclude this paper in Sect. 6.

2 Data Collection

Data collection for this analysis is a tedious and challenging process. Mostly due to security reasons most of the database hack details will be removed by the security crews in no time, so one should be quick in collecting the data and should always be in track of the latest breaches. During the data acquisition, we had to visit some dreadful sites which are not safe and which would have infected our own machine if appropriate security standards were not followed. We even crawled into sites in languages other than English and got the data which made our research even more tough and interesting. We started the breached database collection from 14 Dec, 2013 to 1 Aug, 2014. We have collected data from public databases and any open source information shown in Table 1.

a. *Public databases*: The main source of data collected are pastebin.com, motherboard.vice.com/blog, torrentz.eu, hack the world, lifehacker, exploit-db.com and many other sites.
b. *Open source information*: It includes all blogs and other reviews on the online data breaches. It tells us about the impact and statistics on the corresponding database breach. This information helps in realizing the impact and the security of our online accounts. This also throws light on the vulnerability in the security of some top trusted organizations.

We have also collected large numbers of raw passwords leaked from different websites, such as myspace, rockyou, porn, phpbb, tuscl, singles, alypaa, elitehacker. These are only passwords without usernames; therefore, we have not included them in this research.

Table 1 Collected data details

Website	No. of accounts	Usernames	Passwords	Hashed
Twitter	62,148	Yes	Yes	No
Hotmail	20,000	Yes	Yes	No
Thai4promotion	3000	Yes	Yes	No
Sony	37,608	Yes	Yes	No
Gwaker	188,000	Yes	Yes	No
Snapchat	4.6 million	Yes	No	No
Hi5ads	5000	Yes	Yes	No
Linkedin	6 million	No	Yes	Yes
Just10time	4400	Yes	Yes	Yes
msu.edu	5300	Yes	Yes	Yes
Islamiabank	51	Yes	Yes	Yes
gov.zw.user	590	Yes	Yes	Yes
French site	1280	Yes	Yes	Yes
iCCPS	2770	Yes	Yes	Yes
Powerblog	1000	Yes	Yes	Yes

3 Implementation Details

A Microsoft research [5, 6] found that an average user can remember only 6.5 strong passwords where the user has around 25 online accounts on an average. This motivated us to move forward with our research. Our first experiment was to check the possibility of the password reuse in the collected data.

3.1 Background

In order to find the password reuse, we utilize the usernames and passwords dump of Thai4promotion website which has 3000 username and password combinations. The thai4promotion dump contained the information, such as password, login, name, and email of the user. The possibility that the user might probably be using the same password of the thai4promotion site for logging into his/her attached email accounts was undeniable. Hence, we tried to access their respective email service providers, like Hotmail, Gmail, and Yahoo using the users' thai4promotion website credentials. We found that 20 % of the users were vulnerable to password reuse attack (shown in Fig. 1). This analysis revealed the serious implications of password reuse behavior. Using this 20 % successful email ids and passwords, we checked for the password reuse attack on facebook.com, a worldwide used social networking site. It was more shocking to see that 12 % of these users use same password for Facebook too, which extensively demonstrates the password reuse behavior among users.

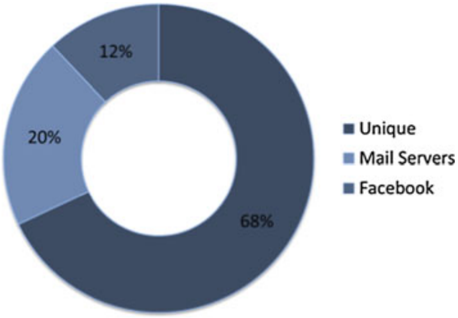


Fig. 1 Password reuse behavior in Thai4promotion website

This above analysis educated us to experiment on password reuse on large number of accounts in order to convey the world about the password reuse behavior and its consequences. So we investigated on the 62,148 usernames and passwords dump leaked from Twitter website. To efficiently carry on the password reuse attack on this huge data we used a sophisticated automation tool, selenium. We leveraged the simplicity and power of python to first arrange the 62,148 nonuniform, leaked usernames and passwords from Twitter website and store them into a file.

3.2 Validation of a Username and Password Pair

In order to check the validity of username/password combinations, we leveraged the power of selenium, a web automation framework for python. At first the tool tries to establish a connection with the target domain and pulls up their login page. Once the page has been completely loaded, the tool fetches up the username/password combinations one-by-one from the already arranged file F mentioned above. For each of the combinations fetched the tool checks for a login. On successful login the tool actually writes the success combination on to another file as well as takes a screen shot of the logged in session to further assert that a login actually happened for further reference. In case any login error occurs, we generate similarly looking passwords for the same username and try to validate the same in a loop till a threshold value *t*.

Similar Password Generator: People not only use the same password but also similar looking passwords for different accounts. For example, a user could keep password “nature” for an account and “n@tur3” or “nature123” for the other. Hence, whenever a login failure event occurs, the password is sent to this module. Based on the Damerau–Levenstein edit distance algorithm we generate semantically similar password for the same username. We perform addition, deletion, transposition, and substitution of characters in the old password using the formula:

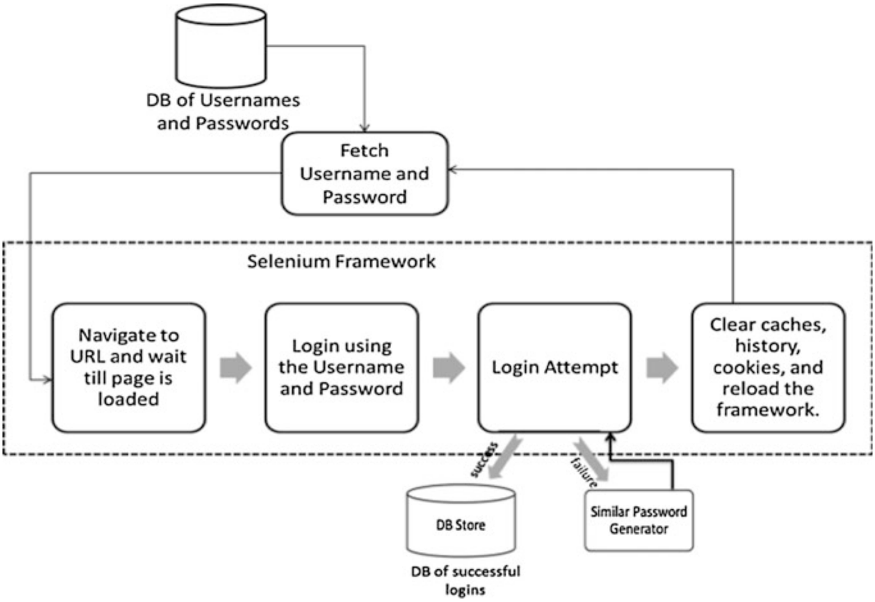

DB of Usernames and Passwords
Fetch Username and Password
Selenium Framework
Navigate to URL and wait till page is loaded
Login using the Username and Password
Login Attempt
Clear caches, history, cookies, and reload the framework.
success
failure
DB Store
DB of successful logins
Similar Password Generator

Fig. 2 Flow diagram of the validation of a username and password

$$\mathrm{DL}\,(\mathrm{old\_password},\,\mathrm{new\_password}) = \min(i)\,[\#\mathrm{Sub}(i) * \mathrm{ws} + \#\mathrm{Del}(i) * \mathrm{wd} + \#\mathrm{Ins}(i) * \mathrm{wi}\;\#\mathrm{TRP}(i) * \mathrm{wt}]$$

where, #Sub, #Del, #Ins, and #TRP are the number of substitutions, deletions, insertions, and transpositions required in the session *i* and ws, wd, ws, and wt are positive numbered weighting factors.

The tool actually has built-in mechanism to ensure that the server does not prompt for a "captcha" challenge. Figure 2 provides the flow diagram of the entire validation process.

4 Experimental Results

We were able to learn that only 63 % of the accounts were valid from the Twitter dump collected. From these valid username/password combinations we tried the password reuse attacks on Facebook, Gmail, Yahoo, and Hotmail with the use of our automation tool. Password reuse attacks on these sites really revealed the bad practice of password reuse and its vulnerability. Our experiments revealed a 33 % password reuse behavior on Facebook, 26 % reuse behavior on Hotmail, 15 % on Gmail and 12 % on Yahoo. Figure 3 shows the reuse behavior among Twitter and Facebook, Gmail, Hotmail and Yahoo.

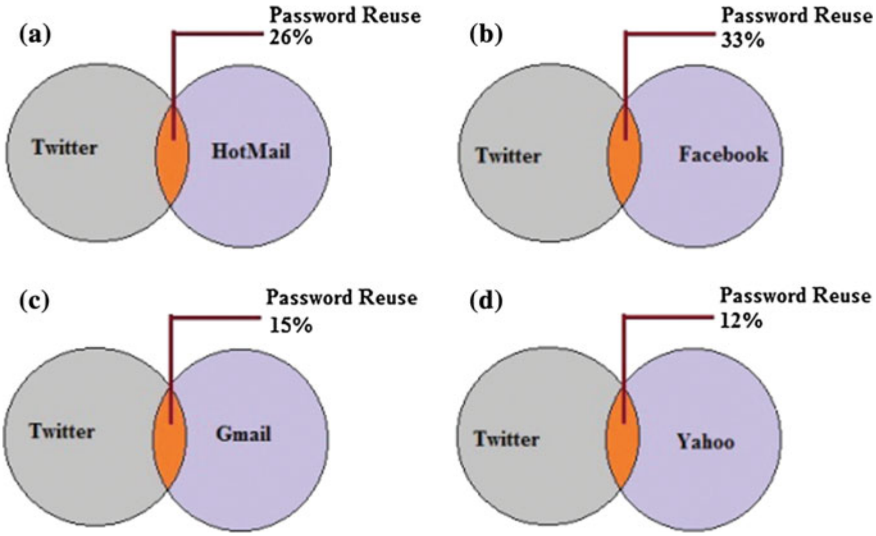
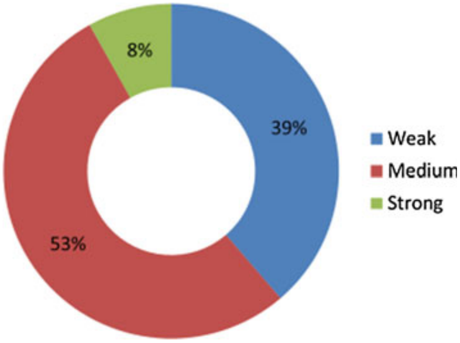


Fig. 3 Password reuse in **a** Hotmail. **b** Facebook. **c** Gmail. **d** Yahoo

Fig. 4 Results of password categorization based on their strengths



By our analysis, password reuse behavior among users is extensively demonstrated and it shows the vulnerability created in the users' account if same or similar password is used across multiple sites, making them prone to get hacked.

Once all set of successful username and password pairs were collected we ran all the passwords through multiple password strength checkers openly available. We tested the strength of the passwords using Microsoft, The Password Meter and Rumkin password strength checker tools. Apart from this we ran all the passwords through our own password strength checker too. Based on the outcomes from different tools we categorized the passwords into three basic categories: weak, medium, and strong as shown in Fig. 4

Our findings show that even if a user has kept a strong password, reusing it or using a similar password in multiple accounts makes it vulnerable.

4.1 Cracking Password Hashes Employing Offline Attacks

To crack the hashed passwords from the collected dumps we used readily available, advanced password recovery tool like hashcat [31]. In order to perform password recovery, the 6 million Linkedin hash which use SHA-1 encryption algorithm was chosen. We were able to crack around 1.5 lakhs of hashes in just 13 min. We experimented on a leaked username/password dump which contains password hashes along with the plain text usernames. We collected credentials from 40dni.sk a Slovakian website. Out of the 4200 hashed passwords collected, we were able to crack only 2048 password. This extensively shows the hidden dreadful side of offline attacks as there is no time barrier to stop the adversary. Many of these username/password credentials are freely available in the internet, once if these are into wrong hands a very dreadful and harmful security treats will be introduced.

5 Survey

To study the internet users' behavior and thinking processes while using many accounts, we conducted a survey of users at different educational universities, including students and professional staffs and others. We received a total of 1508 responses. In the survey, participants had to answer several questions regarding their passwords. Survey was made to demonstrate about the password reuse behavior among the users, which is dangerous.

The major series of questions which the participants were requested to answer in the survey were;

(1) Profession.
(2) Age group.
(3) Number of online accounts.
(4) Number of passwords that the user uses.
(5) Whether the user use a combination of alphanumeric and special characters in their passwords.
(6) Do the passwords have similarity (ex: darknight, DarkKnight, darkKnight1).
(7) Is it difficult for the user to remember the large number of complex passwords?
(8) Do the participants practice password reuse?

These were sent across various educational institutes with a message illuminating the problems of password reuse. A sample of the survey form can be viewed here [32].

5.1 Survey Responses

The survey responses were of great importance as it shows the trend and behavior of the current internet users. The survey responses (as shown in Fig. 5) explain that the participants are from different backgrounds; students, business professionals, professionals, researchers, teachers, and majority of the participants are students between the age categories 18–25. It was found that the majority of the participants have 5–10 online accounts, whereas most of them have 4–8 different passwords.

Only 66 % of the participants (as shown in Fig. 6) use uppercase, lowercase and special characters in all their passwords and around 57 % of the candidates use modified passwords or passwords with similarities (ex: darknight, DarkKnight, darkKnight1). Similar passwords can be easily cracked by brute forcing and dictionary attacks. The core idea of the survey was to infer about the difficulty in remembering long different complex passwords and 71 % of the users find it difficult to remember passwords. Finally, it was found that 59 % of the participants reuse their password across multiple sites because remembering large complex password is difficult and users are forced to reuse passwords.

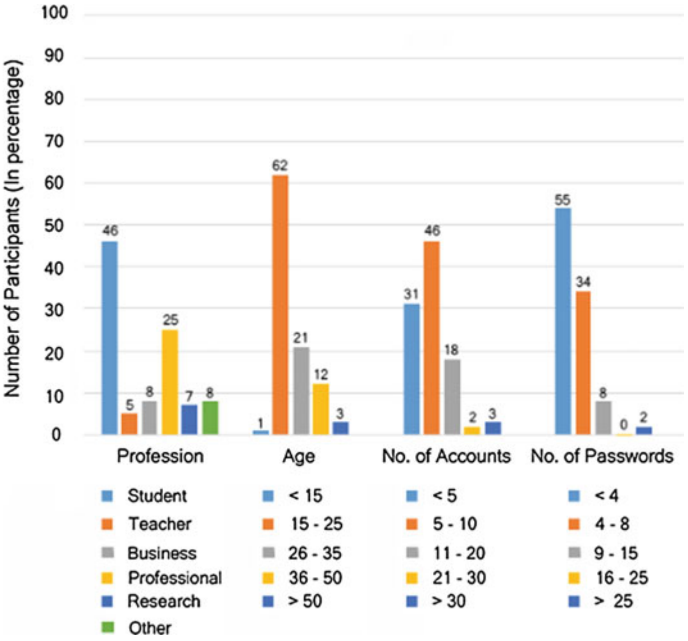


Fig. 5 Survey results of user details

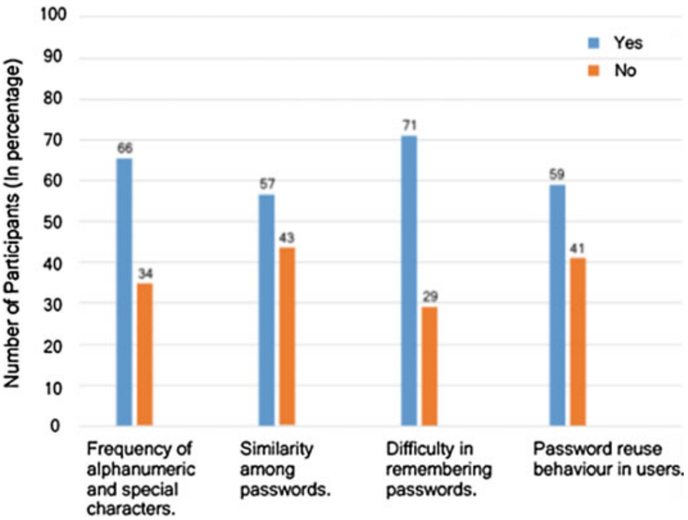


Fig. 6 Survey results of the password reuse behavior among user

6 Conclusions

In this paper, our study shows how password reuse initiates security vulnerability. The act of password reuse could hamper users personal data not from one but a chain of his other accounts all linked with one common password.

We also understood that online data breach even though it could hamper to one company, but password reuse can bring down the personal knowledge management of the users' other connected accounts. We also found that common users find it difficult to remember different passwords for their various accounts which should trigger better research in the field of password management. Also, this study brings in limelight the need of alternate ways for user authentication rather than the trivial text-based passwords.

Acknowledgments We owe a great thanks to many people who have helped and supported us for this project. We thank our Institution and the entire Cyber Security team for their support without whom the completion of this project would have been a distant reality.

References

1. Florencio, Dinei, and Cormac Herley. "A large-scale study of web password habits." Proceedings of the 16th international conference on World Wide Web. ACM, 2007.
2. "Passwords Re-used by Six out of Ten Consumers." *Techworld*. N.p., n.d. Web. 21 Apr. 2015. http://news.techworld.com/security/3400895/passwords-re-used-by-six-out-of-ten-consumers/.

3. Wiedenbeck, Susan, et al. “Design and evaluation of a shoulder-surfing resistant graphical password scheme.” Proceedings of the working conference on Advanced visual interfaces. ACM, 2006.
4. R. Dhamija, J. D. Tygar, and M. Hearst, “Why phishing works,” in CHI ’06: Proc. SIGCHI Conf. Human Factors Computing Systems, New York, 2006, pp. 581–590, ACM.
5. “World’s Biggest Data Breaches & Hacks - Information Is Beautiful.”*Information Is Beautiful Worlds Biggest Data Breaches Hacks Comments*. N.p., n.d. Web. 21 Apr. 2015. http://www.informationisbeautiful.net/visualizations/worlds-biggest-data-breaches-hacks/.
6. “Data Breach and Attacks on Organisations.” N.p., n.d. Web. 21 Apr. 2015. http://www-935.ibm.com/services/us/en/it-services/data-breach/data-breach-statistics.html.
7. Ives, Blake, Kenneth R. Walsh, and Helmut Schneider. “The domino effect of password reuse.” Communications of the ACM 47.4 (2004): 75–78.
8. “Reusing Passwords at Different Websites.” N.p., n.d. Web. 21 Apr. 2015. http://www.researchgate.net/publication/27296513_The_Domino_Effect_of_Password_Reuse.
9. Sun, Hung-Min, Yao-Hsin Chen, and Yue-Hsun Lin. “oPass: A user authentication protocol resistant to password stealing and password reuse attacks.” Information Forensics and Security, IEEE Transactions on 7.2 (2012): 651–663.
10. Devi, S. Megala, and M. Geetha. “OPass: Attractive Presentation of User Authentication Protocol with Resist to Password Reuse Attacks.” (2013).
11. Weir, Matt, et al. “Password cracking using probabilistic context-free grammars.” Security and Privacy, 2009 30th IEEE Symposium on. IEEE, 2009.
12. Narayanan, Arvind, and VitalyShmatikov. “Fast dictionary attacks on passwords using time-space tradeoff.” Proceedings of the 12th ACM conference on Computer and communications security. ACM, 2005.
13. Pinkas, Benny, and Tomas Sander. “Securing passwords against dictionary attacks.” Proceedings of the 9th ACM conference on Computer and communications security. ACM, 2002.
14. Goodin, Dan. ““thereisnofatebutwhat-wemake”-Turbo-charged Cracking Comes to Long Passwords.” Ars Technica. N.p., n.d. Web. 21 Apr. 2015. http://arstechnica.com/security/2013/08/thereisnofatebutwhatwemake-turbo-charged-cracking-comes-to-long-passwords/.
15. “Preventing Password Reuse.” *Preventing Password Reuse*. N.p., n.d. Web. 21 Apr. 2015. http://www.slyman.org/blog/2011/02/preventing-password-reuse/.
16. A Study Of Password Habits Among American Consumers, and September 2012. *CONSUMER SURVEY: PASSWORD HABITS* (n.d.): n. pag. *Against Fraud Attacks*. Web. 21 Apr. 2015. http://www.csid.com/wp-content/uploads/2012/09/CS_PasswordSurvey_FullReport_FINAL.pdf.
17. J. Bonneau, “The science of guessing: analyzing an anonymized corpus of 70 million passwords,” in Proceedings of the 33rd IEEE Symposium on Security and Privacy, ser. SP ’12, May 2012.
18. Kelley, Patrick Gage, et al. “Guess again (and again and again): Measuring password strength by simulating password-cracking algorithms.” Security and Privacy (SP), 2012 IEEE Symposium on. IEEE, 2012.
19. Ms. A. G. Khairnar and Prof. N. L. Bhale. “A Survey on Password Security Systems.” IJECSE, Volume2,Number 2, April 2013.
20. Gaw, Shirley, and Edward W. Felten. “Password management strategies for online accounts.” Proceedings of the second symposium on Usable privacy and security. ACM, 2006.
21. Komanduri, Saranga, et al. “Of passwords and people: measuring the effect of password-composition policies.” Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, 2011.
22. “Krebs on Security.” *Krebs on Security RSS*. N.p., n.d. Web. 21 Apr. 2015. http://krebsonsecurity.com/2013/11/cupid-media-hack-exposed-42m-passwords/comment-page-1/.
23. “Sony Password Analysis.” *Sony Password Analysis*. N.p., n.d. Web. 1 Jan. 2015. http://www.troyhunt.com/2011/06/brief-sony-password-analysis.html.

24. “Lulzsec’s Sony Hack Shows Rampant Password Reuse.” *Lulzsec’s Sony Hack Shows Rampant Password Reuse*. N.p., n.d. Web. 21 Apr. 2015. http://www.computerworld.com/s/article/9217646/LulzSec_s_Sony_hack_shows_rampant_password_re_use.
25. “Sony Hack Reveals Password Security Is Even Worse than Feared.” • *The Register*. N.p., n.d. Web. 21 Apr. 2015. http://www.theregister.co.uk/2011/06/08/password_re_use_survey/.
26. J. A. Halderman, B. Waters, and E. W. Felten, “A convenient method for securely managing passwords,” in WWW ’05: Proc. 14th Int. Conf World Wide Web, New York, 2005, pp. 471–479, ACM.
27. Egelman, Serge, et al. “It’s Not Stealing If You Need It: A Panel on the Ethics of Performing Research Using Public Data of Illicit Origin.” Financial Cryptography and Data Security. Springer Berlin Heidelberg, 2012.
28. Zhang, Yinqian, Fabian Monrose, and Michael K. Reiter. “The security of modern password expiration: an algorithmic framework and empirical analysis.” Proceedings of the 17th ACM conference on Computer and communications security. ACM, 2010.
29. “Most Common and Hackable Passwords on the Internet.” *Most Common and Hackable Passwords on the Internet*. N.p., n.d. Web. 21 Apr. 2015.http://www.telegraph.co.uk/technology/internet-security/10303159/Most-common-and-hackable-passwords-on-the-internet.html.
30. “Beware-meta-password Reuse.” *Beware-meta-password Reuse*. N.p., n.d. Web. 23 Mar. 2015. http://www.itworld.com/tech-society/54193/beware-meta-password-reuse.
31. Steube, J. “Hashcat Advanced Password Recovery.” (2013).
32. “Survey : Impact Of Massive Online Breaches On Password Reuse Behaviour.” *Google Docs*. N.p., n.d. Web. 21 Apr. 2015. https://docs.google.com/forms/d/1Ig8GFGC0rry7gwWOC-BMuvSS2NBy3X3Zl1J7TGbONu4s/viewform.

Identifying Significant Features to Improve Crowd Funded Projects' Success

Jaya Gera and Harmeet Kaur

Abstract As volume of projects launched on various crowdfunding website is growing with time, success percentage is decreasing. So, it is important to understand project dynamics and its chance to succeed. Projects launched over crowdfunding platform vary in nature, aim and quality and so is their probability of success. Recent research indicates that various project features, such as, size of social networks, pledging behavior, timing of investments, accumulated capital, etc. influence success of a project. In this paper, significant project features that should be improved upon in order to promote campaigns are identified. It is used in conjunction with the pledging behavior of different projects so that the projects that are likely to be successful can be timely promoted.

Keywords Crowdfunding · Pledge behavior · Funding cycle · Campaigns success · Significant features

1 Introduction

Crowdfunding has attracted a number of new ventures to meet their financial needs for setting up start-ups business because it provides a simple and easy access to funds than the traditional fundraising methods. Crowdfunding allows entrepreneur who request resources, to appeal for funds directly from investors who provide resources, through online platforms [1].

J. Gera (✉)
Department of Computer Science, Shyama Prasad Mukherji College,
University of Delhi, Delhi, India
e-mail: jayagera@spm.du.ac.in

H. Kaur
Department of Computer Science, Hans Raj College,
University of Delhi, Delhi, India
e-mail: hkaur@hrc.du.ac.in

H.S. Saini et al. (eds.), *Innovations in Computer Science and Engineering*,
Advances in Intelligent Systems and Computing 413,
DOI 10.1007/978-981-10-0419-3_25

Crowdfunding is small in terms of overall economic activity, but it is growing in terms of volume and dimension [2]. Crowd funding initially started with music industry, has now been expanded to social, political, technical, sports, education, finance, investment, and many more [3]. Crowdfunding platforms have grown like mushrooms in a decade's time. Some of the well-known platforms are Kickstarter, IndieGoGo, Sellaband, and DonorChoose [4]. Thousands of projects are launched on these websites, but not all are successful to raise sufficient funds. Every project launched on a crowdfunding site has a deadline and a funding goal (target amount) to raise. Success here means that project achieves its funding goal.

A major issue is to understand the factors working behind success. Success of a campaign depends on various factors such as goal amount, its duration, quality, project presentation, funding behavior, etc. Chances of success of a project can be improved at pre-launch as well as at post launch stage. Pre-launch success probability depends on project preparation and quality. Project representation and preparation is positively associated with project success. Crowdfunding projects that signal high quality are more likely to be funded [5]. Project representation includes project video, project description, goal amount, duration, etc. These features should be weighed and improved prelaunch to improve chances of success of a project. Postlaunch a project's success can be improved by working with funding pattern [6–8] and improving features such as projects updates and number of rewards.

Contributions of the work presented in the paper are as follows:

1. To identify the features those are significant for the success of crowd funded projects.
2. To increase the probability of success of the projects.
3. Which projects should be promoted and when?

Rest of the paper is organized as follows: Sect. 2 investigates related work. Section 3 describes dataset and its main characteristics. Section 4 explains factors that are prominent in influencing success of a crowd funded project. Section 5 presents analysis of pledge behavior. Section 6 concludes the work and pays attention to future work and the next section is for acknowledgement.

2 Literature Survey

Crowdfunding is an emerging phenomenon that has recently attracted many academicians and research scientists. Much of the research has been focused on market, economic, and investment perspective of crowdfunding. A number of the literatures are available in the form of articles and working paper, but it lacks in scientific and peer reviewed work [5].

Lambert et al. [9] is the first one to investigate crowdfunding projects characteristics and drivers of success. Analysis indicated that nonprofit organizations are more likely to be successful in raising funds than any other form of organizations. Greenberg et al. [10] predicted success of crowdfunding campaign on the basis of various static project features. Mollick [5] also studied the dynamics of success and failure of crowdfunding campaigns. This study revealed that project preparation and size of social network influence project success. Project quality plays an important role in increasing chances of success. Etter et al. [6] too predicted success of a crowdfunding campaign using information such as percentage of funds raised, number of backers and number of twits made during funding cycle. Agarwal et al. [2] suggested that crowdfunding overcomes distance-related economic frictions. Geographical distance between entrepreneur and investor plays reduced role. Local investors tend to invest relatively early and are not much influenced by other funders' decision. Ordanini et al. [11] examined the role of consumer in crowdfunding market and how consumers turn to investors. Giudici et al. [12] examined influence of social capital on success. Zvilichovsky et al. [13] analyzed the creator's backing history and its impact on success of a project and concluded that owners' who are actively involved in backing other projects attract more backers and are more likely to be successful. Mitra et al. [14] correlated success of a campaign and words/phrases used in project descriptions on a crowdfunding platform.

In the literature, the focus is to improve the success probability by identifying significant project features that should be taken into consideration at the launch time. Many project features are static and can not be improved, once a project is launched. So, this study also analyses funding pattern to identify projects that are lagging behind and should be promoted timely to cover up the lost ground. In this paper, significant project features that are dynamic such as project updates and should be paid attention to improve probability of a project's success during the funding cycle are also identified.

3 Dataset Description

The dataset used is released by the owner of web site: http://www.kickspy.com/. This was a web site that used to scrap data from Kickstarter and to analyze projects and to provide useful statistics on projects. This website has been shut down. This dataset contains data scraped of Kickstarter projects launched in the month of April 2014. This dataset contains project features, pledged money, and rewards offered to funders of 4682 projects. Project features include project id, name of project, target amount, pledged amount, status, category, number of updates, number of comments posted, number of backers, start date, end date, duration, etc. Pledge data consists of amount pledged during funding cycles by these projects. Reward data consists of number of different level of rewards offered to backers.

We have used project features and pledge data for purpose of analysis. Project data consists of number of project features. Project features that are of key interests for analysis are listed below:

1. Top_Category: Category of project to which it belongs
2. Goal: Target Amount to be raised
3. Has_video: Project page on Kickstarter site has video or not

4. Rewards: Number of rewards offered by creator
5. Duration_in_Days: Number of days for which project remain live on crowdfunding platform and available for funding purposes
6. Facebook_Connected: Project has link to Facebook account (Yes or No)
7. Facebook_Friends: Number of friends creator has on Facebook
8. Facebook_Shares: How many have shared projects on Facebook
9. No_of_Images: Number of Images uploaded on the web site
10. No_of_Words_Description: Number of words used in project description
11. No_Words_Risks and Challenges: Number of words used in describing risk and challenges of project
12. Creator_Projects_Created: Number of crowdfunding projects created by Creator
13. Creator_Projects_Backed: Number of crowdfunding projects backed by Creator
14. Update: Number of updates uploaded on website
15. Comments: Number of comments
16. Status: Project status—successful or unsuccessful
17. Pledged Amount: Total Amount raised
18. Number of Backers: Total number of backers backed the project

Pledge data consists of project id, date of pledge, amount pledged, and number of backers. This dataset consists of data about successful (1900 projects), unsuccessful (2235 projects), canceled (482 projects), suspended (6 projects), and live projects (59 projects). Data of successful and unsuccessful projects are used for analysis purpose.

4 Prominent Factors

Our aim is to improve success probability of a project launched over a crowd funding platform. Project success probability can be improved prelaunch as well as postlaunch. Though, crowdfunding seems to be an easily approachable option to a venture. Raising funds successfully requires variety of skills and lot of efforts by creator [15]. Creator needs to understand the responsibility, market need, funders' expectation, associated risk, and importance of timing of events. A project success is determined by project quality and project preparation. A creator can work on project quality and promote project locally to improve success of project prelaunch.

Project success is also determined by various project features. To determine project features that are more dominant, we performed attribute evaluation using method—WrapperSubsetEval with three different classifiers—Classification by regression, logistic regression, and J48. This experiment is performed only for successful and unsuccessful projects and on first 16 attributes listed in Sect. 3. Pledged amount and number of backers are not considered because they are not known at the time of launch. Although, updates and comments are also not known at the time of launch, they are included in the experiment, as number of updates

Table 1 Attributes selected by attribute evaluator—WrapperSubsetEval

Classifier—classification by regression	Classifier—SimpleLogistic	Classifier-J 48
Top_Category Updates Comments Goal Duration_in_Days Facebook_Connected Facebook_Shares Creator_Projects_Created Creaor_Projects_Backed	Top_Category Updates Comments Goal Duration_in_Days Facebook_Friends Facebook_Shares Has_Video	Comments Goal Facebook_Connected Facebook_Shares

(and sometimes comments also) are posted by the creator after launch and influences success of a campaign [16]. Also, these factors are under the control of the creator and the creator can work on them during funding cycle so that the project gets successfully funded. Table 1 shows the significant features as per three different classifiers used with WrapperSubsetEval. Out of 16 attributes that were considered, different classifiers found different sets of significant attributes.

From Table 1 it can be concluded that Top_Category, Updates Comments, Goal, Duration_in_Days, Facebook_Connected and Facebook_Shares are the attributes that play more important role in successful funding of a project than the other features. Out of these attributes, Top_Category, Goal and Duration_in_Days are the ones that cannot be changed after the launch of the project. So, the creator should be careful about them at the launch of the project. The rest of the attributes gain importance during the funding cycle of the project. These are the features on which focus should be if a project is lagging during the funding cycle as explained in the later part of this section.

We also executed various machine learning algorithms on the dataset with these 16 attributes to predict probability of success of projects and the results are shown in Table 2. The RandomTree and NaiveBayes does not give the ordering of features so they are not shown in Table 1.

Table 2 Accuracy of classifier for the given dataset

Classifier	Accuracy (%)
Classification by regression	83.7929
Simple logistic	83.0189
J48	82.1239
Random tree	76.5119
Naïve Bayes	66.9569

5 Pledge Behavior

A campaign to be improved is identified not only using project features at the time of launch or prelaunch but also using its raised amount during funding cycle. If a project is already launched, then there is little scope to improve some features, such as category, goal amount, duration, quality of video, images, etc. But, there are features that can be improved, such as number of rewards, level of rewards, product, or services features to be delivered, etc.

Project success can also be monitored using pledge behavior analysis and projects that have already generated some percentage of funding can be considered. But how much a project should raise to come under this clause and by what time?

This dataset under consideration contains data about successful, unsuccessful, canceled, live, and suspended projects. This study focuses on understanding nature of pledge behavior of successful and unsuccessful campaigns. So, for this analysis we do not consider other such as canceled campaigns.

Dataset is resampled and normalized to perform pledge behavior analysis. The time and amount of pledge money is resampled to have uniform number of samples for all campaigns. Pledge money amount is divided by goal amount to obtain percentage of goal amount pledged. Successful projects and unsuccessful projects have different funding behavior pattern. Successful projects attract more funds than unsuccessful ones and growth pattern of successful projects are higher than unsuccessful ones. Significant difference is observed in the funding pattern of successful and unsuccessful projects soon after their launch and this gap between funding ratio keeps on increasing. We are interested in knowing what pattern a project follows to become successfully funded. So, we studied the funding behavior pattern of projects that are funded approximately 100 % by plotting median of funds received.

Figure 1 shows growth pattern of projects received funds between 100 and 108 %. If a campaign has raised approximately 20 % of its goal within the first 15 % of funding cycle, its success probability is high. Campaigns that could raise 20 % of funds within 20 % of funding time should be identified and should be promoted.

After the launch of a project, if it is observed that a campaign has achieved 20 % funding within 20 % of the funding cycle, then these projects should be promoted by working at three levels:

- By improving project features that have been identified as significant in Sect. 3.
- By publicizing these projects over social media via Facebook, twitter etc.
- By recommending funders to invest in these projects.

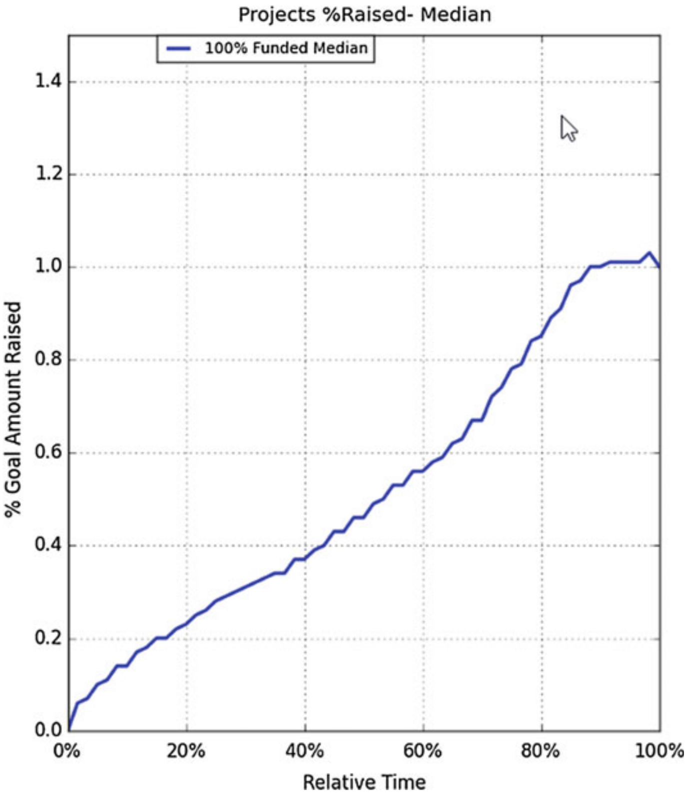


Fig. 1 Pledge money graph

6 Conclusion

Achieving target amount of a campaign is crucial milestone for a new venture. Finding earlier the performance of a campaign can help in meeting target later in funding stage. In this paper, we analyzed pledge behavior of different campaigns and identified campaigns to be promoted and improved on various factors. We also studied various factors influencing project success and their significance. To improve a project's success, creator, platform, and funders need to play their role carefully. In our future work, we will understand how to promote such campaigns among funders. Potential investors matching with project profile will be identified and will be recommended to invest in such projects.

Acknowledgments This analysis is based on dataset downloaded from web site: http://www.kickspy.com/. This is released by owner of this website. This dataset is rich in Kickstarter projects features and enabled us to glance into the characteristics of projects and its association with success. We would like to thank owner of this website for uploading this dataset and allowing others to use it for research work.

References

1. Gerber, E.M., Hui, J.: Crowdfunding: Motivations and Deterrents for Participation. In: ACM Transactions on Computer-Human Interaction (TOCHI) 20, vol. 20, Issue 6, article no 34, 34:1–34:32, New York, NY, USA (2013).
2. Agrawal, A.K., Catalini, C., Goldfarb, A.: The Geography of Crowdfunding. NBER working paper 16820 available at:http://www.nber.org/papers/w16820 working paper (2011).
3. Hemer J.: Snapshot on Crowdfunding. Working Paper R2/2011. Fraunhofer Institute (2011)
4. Wash, R., Solomon, J.: Coordinating Donors on Crowdfunding Websites. In: The ACM's Conference on Computer Supported Cooperative Work and Social Computing, Baltimore, Maryland, USA (2014).
5. Mollick, E.: The dynamics of crowdfunding: An exploratory study. Journal of Business Venturing 29, 1–16 (2014).
6. Etter, V., Grossglauser, M., Thiran, P.: Launch Hard or Go Home!: Predicting the Success of Kickstarter Campaigns. In: The Proceedings of the First ACM Conference on Online Social Networks, pp 177–182, New York, NY, USA (2013).
7. Wash, R.: The value of completing crowdfunding projects. In: Emre Kiciman; Nicole B. Ellison; Bernie Hogan; Paul Resnick & Ian Soboroff, ed., 'ICWSM', The AAAI Press, Boston, Massachusetts, USA (2013).
8. Burtch G., Ghose A., Wattal S.: An empirical examination of the antecedents and consequences of investment patterns in crowd-funded markets. In: SSRN Electronic Journal (2011).
9. Lambert T., Schwienbacher A.: An Empirical Analysis of crowdfunding. Mimeo.Louvain School of Management, Belgium. Available at SSRN: http://ssrn.com/abstract=1578175 (2010).
10. Greenberg, M.D., Pardo, B., Hariharan, K., Gerber, E.: Crowdfunding Support Tools: Predicting Success & Failure. In: The CHI 2013 Extended Abstracts on Human Factors in Computing Systems, pp 1815–1820, New York, NY, USA (2013).
11. Ordanini A., Miceli L.,Pizzetti M. Parasuraman A.: Crowdfunding: transforming customers into investors through innovative service platforms. In: Journal of Service Management, Vol. 22 Issue 4 pp. 443–470 (2011).
12. Giudici G., Guerini M. and Lamastra C. R.: Why crowd funding can succeed: role of proponents' individual and social capital. In: SSRN Electronic Journal, doi:10.2139/ssrn.2255944, (2013).
13. Zvilichovsky D. Inber Y., Barzilay O.: Playing both sided of the market: Success and reciprocity on crowdfunding platforms. Working paper, Tel Aviv University, Tel Aviv, Israel (2013).
14. Mitra T., Gilbert E,: The language that gets people to give: Phrases that predict success on Kickstarter. In: CSCW '14, ACM, pp 49–61 New York, NY, USA (2014).
15. Hui, J.S., Gerber, E., Greenberg, M.: Easy Money? The Demands of Crowdfunding Work. Segal Technical Report. 12–04 (2012).
16. Xu, A., Yang, X., Rao, H., Huang, S.W., Fu, W.-T., Bailey, B.P.: Show me the Money! An Analysis of Project Updates during Crowdfunding Campaigns. In: The Proceedings of the SIGCHI Conference on Human Factors in Computing, pp 591–600, New York, NY, USA (2014).

A Survey on Challenges in Software Development During the Adoption of Agile Environment

Chandrika Sikka, Saru Dhir and Madhurima Hooda

Abstract In the recent years, agile development has received increasing interest in software development organization. Agile adoption has lead to increase in productivity and has better outcome than the traditional software development methods. However, agile adoption also come with challenges which can adversely affect the project. In this paper, the challenges of agile implementation and its solutions are discussed.

Keywords Agile methodology · Methods · Challenges · Success and failure rate · Complexities

1 Introduction

Software development life cycle (SDLC) is the process for planning, creating, testing, and deploying the software products. Software engineers use software development models for building error free software that meets customer requirements so as to deliver the software within the specified time limit. Traditional models give the improper division of software into many stages. The disadvantage of waterfall model is that the requirements are fixed at the beginning and it is difficult to react to further changes [1].

Agile software development consists of methodologies that develop the software using iterative and incremental method. The focus of agile software development is different from traditional software development in a way that it provides the

C. Sikka (✉) · S. Dhir · M. Hooda
Amity University, Noida, Uttar Pradesh, India
e-mail: chandrikasikka5@yahoo.com

S. Dhir
e-mail: sdhir@amity.edu

M. Hooda
e-mail: mhooda@amity.edu

H.S. Saini et al. (eds.), *Innovations in Computer Science and Engineering*,
Advances in Intelligent Systems and Computing 413,
DOI 10.1007/978-981-10-0419-3_26

interaction between the customer and developer. The customer feedback helps to improve the software with an updates and hence fast delivering of software.

In spite of the benefits of agile software development, there are many challenges faced by distributed agile development process. This paper is organized as follows: Sect. 2 gives the literature review of agile software development. Section 3 presents the different agile methodologies. Section 4 describes agile challenges and its solutions. Section 5 describes the success and failure rate on the basis of technical, domain and organizational complexity. Finally, Sect. 6 presents the conclusion.

2 Literature Review

The literature review on agile focused on challenges while implementing it. The related work was done by various people. Some of them are:

Many papers discuss case studies and their corresponding solutions [2]. According to the June 2012 survey report, it was found that 71 % of organizations work done agile and 15 % tried to switch on agile [3]. In January 2014 survey report, the agile project had a success rate of 33 % and instead of that 11 % agile projects achieve great success [4].

An approach was developed to deliver the product on time in comparison to traditional SDLC. Different steps were defined to adopt the development methods in the organization and also a case study was discussed for the adoption of this new approach [5].

3 Agile Methodologies

In IT field agile software development methods have gained much popularity. The methodologies such as: Scrum, eXtreme programming, lean, feature-driven development [6, 7].

3.1 Agile Scrum Methodology

In this methodology product development occurs in small pieces and building products one small piece at a time. The scrum consists of product owner, development team, scrum master, stakeholders, and managers. The Scrum process is suited for projects where the requirements are changing rapidly. Scrum software development contains the series of iterations called sprints. Each sprint starts with a brief planning meeting and ends with a review.

3.2 Agile eXtreme Programming

Extreme Programming (XP) defines as a software development methodology based on values of simplicity, communication, feedback, respect and courage used to improve software quality. XP teams follow a plan by deciding what should be done next and predict when the project will be done. Instead of delivering the whole software, this process delivers the software as we need it. The pair programming is done in this method and work all together. The Extreme Programming mainly stresses on customer satisfaction.

3.3 Lean Methodology

Lean Development is an iterative agile methodology. Lean Development mainly focuses on delivering value to the customer. The principles [8] of Lean methodology includes eliminating waste, amplifying learning, deciding as late as possible, delivering as fast as possible, empowering the team, building integrity in and seeing the whole.

3.4 Feature-Driven Development (FDD)

Feature-driven development (FDD) is defined as an iterative and incremental software development process. It consists of five different activities, such as develop overall model, build feature list, plan by feature, design by feature, and build by feature.

3.5 Dynamic Systems Development Method (DSDM)

DSDM is also defined as an iterative agile development model which is a combination of people's knowledge, tools and techniques based on best practices to deliver new systems. The goal of DSDM is to deliver working systems in short period of time.

4 Challenges and Solutions in Agile Development

On developing software with agile methodology organization have faced many problems. Table 1 represents different parameters those having different challenges and its solutions in agile environment.

Table 1 Different challenges and its solutions in agile environment

S. no	Parameters	Challenges	Solutions
1	Responsibilities	Team members did not that they collectively responsible for the overall development [8]. They consider themselves to be responsible for their own part	To overcome this problem team should follow hierarchical structure
2	Directness and honesty	The communication done by offshore and onshore teams are not always fruitful. The offshore team shows only positive points to the onshore team	They should be provided with third party, which leads to a more open and free communication [12]
3	Language barrier	Both offshore and onshore teams faced a problem in communication due to different languages spoken all over the world	The remedy for this is to conduct language classes [13]
4	Distant collaboration	The distributed nature of team became a challenge to find efficient ways of communication	Phone calls, email and messages are the only mode of communication
5	Skill differences	There must be difficulties in technical skills b/w offshore and onshore teams	Write little documentation so as to maintain communication
6	Technical issues	Difficulties in the availability of technology and infrastructure may result in the lack of collaboration b/w offshore and onshore teams	The solution for this problem is by providing an initial training by an agile coach [7, 13]
7	Increased documentation	Due to lack of close collaboration, it would not be possible to maintain less documentary work	Support of developers received by offshore teams in order to avoid gathering of technical debt, and the offshore teams would be given less documented task [14]
8	Sprint planning meeting	Sprint planning meeting may not be possible due to cultural differences, time zone or other holidays	With the help of linguistic experts communication can be easily done

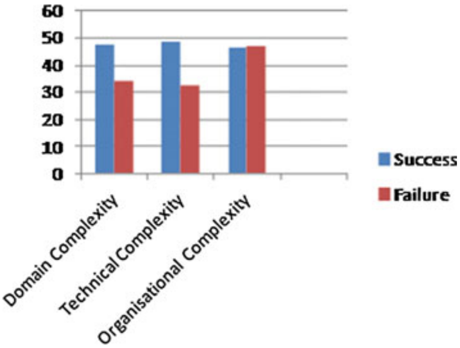
5 Success and Failure Rate on the Basis of Technical, Domain, and Organizational Complexity

As per the year 2012 survey, it was analysed that there are few factors responsible to adopt agile methodology such as managing changing priorities which is 90 %, increase productivity up to 85 %, enhance software quality by 81 %, reduce risk by

80 %, improved engineering disciplines at the rate of 74 % [9]. Many people are noticing that agile development is beneficial to business with an increase of 11 % over last 2 years. In 2013, the reason for adopting agile is that there is an increase in the rate of above-mentioned factors from 90 to 92 % in changing priorities, increase in the rate of productivity by 2 %, enhance software quality up to 82 %, able to reduce risk by 82 % and improved engineering disciplines at the rate of 74 % [10]. The top three benefits of agile development remain: Ability to manage changing priorities (87 %), team productivity (84 %), enhance software quality (78 %), reduce risk (76 %), and improved engineering disciplines (72 %) [11]. One of the scaling factors of the Agile Scaling Model (ASM) is *technical complexity*. The fundamental observation is that the underlying technology of solutions varies and as a result your approach to developing a solution will also need to vary. Factors faced by team are new technology and multiple technology platforms, legacy data and systems, and commercial off-the-shelf (COTS) solutions. Increased *domain complexity* may affect the strategy in many ways such as reaching initial stakeholder consensus becomes difficult, increased prototyping during inception, increased requirements exploration, medium complexity, and high risk. The last complexity is *organizational complexity* where the existing organization may reflect traditional values, through which the complexity of adopting and scaling agile strategies increases within the organization. Factors on which organizational success and failure rate depends are lack of stakeholders and trust, involvement, different visions, and management resistance.

From the above factors it is calculated that the average rate of success and failure in domain, technical, and organizational complexity which is shown in the Fig. 1. It is seen that the factors affecting both domain and technical complexity has high success rate than failure rate. But, due to some factors involve in organizational complexity produce the average rate of failure higher than the success rate.

Fig. 1 Graph showing success and failure rate of three complexities



6 Conclusion

Agile approaches are developed to increase the flexibility and their advantages encourage software companies to use the agile methods, but they also face various challenges by following them. The paper discusses the challenges faced by various organizations while implementing software using agile methodology. The challenges are mainly in management, organizational culture and people, and process area. Paper also discussed respective solutions which may help in improving the challenges faced in agile methodology.

References

1. Preeti Rai, Saru Dhir. Impact of Different Methodologies in Software Devlopment Process in (IJCSIT) International Journal of Computer Science and Information Technologies. Vol. 5(2), ISSN:0975-9646, pp. 1112–1116 (2014).
2. Mira K. Mattsson, Gayane Azizyan, M. K. Magarian, Classes of Distributed agile development problems. in Proc. Agile 2010 conference. IEEE Computer Society. pp. 51–57 (2010).
3. http://www.ambysoft.com/surveys/stateOfITUnion201209.html.
4. http://www.ambysoft.com/surveys/agileJanuary2014.html.
5. Sahil Aggarwal, Saru Dhir. Swift Tack: A New Development Approach in International Conference on issues and Challenges in intelligent Computing Techniques in IEEE (2014).
6. Sahil Aggarwal, Saru Dhir. Ground Axiomsto Achieve Movables: Methodology in International Journal of Computer and Applications Vol-69, No. 14. (2013).
7. Dr. Deepak Kumar, Saru Dhir. A Role of Non-Functional Computing in Software Engineering. ICRITO p. 133 (2013).
8. J.M. Robarts. Practical considerations for distributed projects. in Proc. AGILE 2008 conference. IEEE Computer society Toronto Canada. pp. 327–332 (2008).
9. www.versionone.com/pdf/2012-state-of-agile-survey.
10. www.versionone.com/pdf/2013-state-of-agile-survey.
11. www.versionone.com/pdf/2014-state-of-agile-survey.
12. B. Drummond and JF Unson. Yahoo! distributed Agile: Notes from the world over in Proc. AGILE 2008 Conference Toronto IEEE Computer Society Canada. pp. 315–321 (2008).
13. E. Uy and N. Loannou. Growing and sustaining an offshore Scrum engagement in Proc. AGILE 2008 Conference IEEE Computer Society Toronto. pp. 345–350 (2008).
14. G.M. Cottmeyer. The goods and bads of Agile offshore development in Proc. AGILE 2008 Conference, IEEE Computer Society Toronto. pp. 362–367 (2008).

A Novel Approach for Detection of Motion Vector-Based Video Steganography by AoSO Motion Vector Value

Srinivas Bachu, Y. Olive Priyanka, V. Bhagya Raju and K. Vijaya Lakshmi

Abstract Although tremendous progress has been made on steganography in last decade but still there exist problems to detect the steganalysis in motion-based video where the content is consistently is in motion behavior which creates hurdles to detect it. The motion value plays a crucial role in analyzing the difference between the rated, which allows us to focus on the difference between the actual SAD and the locally optimal SAD after the adding or subtracting one operation on the motion value. Finally, to perform the classification and extraction process-based motion vectors, two feature sets are been used to complete this process based on the fact that most motion vectors are locally optimal for most video codecs. The proposed method succeeds to meet the application requirement and simultaneously succeed in detecting the steganalysis in videos compared to conventional approaches reported in the literature.

Keywords Steganography · AoSO · SAD · SVM · DCT · MV

S. Bachu (✉) · V. Bhagya Raju
Department of ECE, Guru Nanak Institutions Technical Campus,
Ibrahimpatnam, Telangana, India
e-mail: bachusrinivas@gmail.com

V. Bhagya Raju
e-mail: vbhagya01@gmail.com

Y. Olive Priyanka
DECE, Guru Nanak Institutions Technical Campus, Ibrahimpatnam,
Telangana, India
e-mail: olivepriyanka6@gmail.com

K. Vijaya Lakshmi
Department of CSE, Guru Nanak Institutions Technical Campus,
Ibrahimpatnam, Telangana, India
e-mail: vldms@yahoo.com

H.S. Saini et al. (eds.), *Innovations in Computer Science and Engineering*,
Advances in Intelligent Systems and Computing 413,
DOI 10.1007/978-981-10-0419-3_27

1 Introduction

The objective of steganalysis is to identify the vicinity of covertly concealed information in an object. Advanced media documents, for example, pictures, feature, and sound, are perfect spread items for steganography as they commonly comprise countless components that can be somewhat changed to implant a mystery message. Also, such observational spreads are somewhat hard to model precisely utilizing measurable descriptors, which significantly entangles location of installing changes. Specifically, except for a couple of neurotic cases, the recognition cannot be in light of assessments of the basic likelihood disseminations of insights extricated from spread and stego-objects. Rather, identification is generally given a role as an administered grouping issue actualized utilizing machine learning [1, 2].

Albeit there exists a vast assortment of different machine learning apparatuses, bolster vector machines (SVMs) appear to be by a wide margin the most prevalent decision. This is because of the way that SVMs are sponsored by a strong numerical establishment cast inside of the factual learning hypothesis and on the grounds that they are impervious to overtraining and per frames rather well notwithstanding when the component dimensionality is practically identical or bigger than the extent of the preparation set. Additionally, hearty and effective open-source implementations are accessible for download and are anything but difficult to utilize [3].

The unpredictability of SVM preparing, on the other hand, backs off the improvement cycle notwithstanding for issues of a moderate size, as the intricacy of figuring the Gram network speaking to the piece is corresponding to the square of the result of the component dimensionality and the preparation set size. In addition, the preparation itself is in any event quadratic in the quantity of preparing specimens. This force constrains on the extent of the issue one can deal with practically speaking and powers the steganalyst to intentionally plan the elements to fit inside of the multifaceted nature requirements characterized by benefit capable processing assets. Group classifiers give generously more flexibility to the examiners, who can now plan the components essentially without requirements on highlight dimensionality and the preparation set size to assemble locators through a much speedier improvement cycle [4, 5].

Early component-based steganalysis calculations utilized just two or three dozen elements, e.g., 72 higher request snippets of coefficients got by changing a picture utilizing quadratic mirror channel, paired likeness metric, discrete cosine change (DCT) highlights, and higher request snippets of wavelet coefficients. Expanded complexity of steganography calculations together with the longing to identify steganography all the more precisely provoked steganalyst to utilize highlight vectors of progressively higher dimension. The list of capabilities intended for JPEG pictures depicted in utilized elements and was later reached out to twice its size via Cartesian alignment, while 324- and 486-dimensional element vectors were proposed in and separately. The SPAM set for the second request Markov model of pixel contrasts has a dimensionality of 686. Additionally, it demonstrated helpful to consolidation elements processed from distinctive areas to further expand the assorted qualities. The 1234-dimensional cross-space highlight (CDF) set demonstrated particularly viable against YA, which rolls out inserting improvements in a key subordinate area [6, 7].

Besides, such observational spreads are fairly hard to model precisely utilizing statistical descriptors, which considerably confounds discovery of inserting changes. Specifically, except for a couple of neurotic cases, the recognition cannot be in light of appraisals of the hidden likelihood circulations of measurements extricated from spread and stego-objects. Rather, identification is normally giving a role as a regulated classification issue actualized utilizing machine learning [7].

2 Proposed Methodology

2.1 Assumptions for Motion Vector-Based Steganography

(i) The stego noise $\eta_{k,l}$ is assumed to be independent of $V_{k,l}$ and of each other, which is a reasonable assumption if the stego noise is encrypted or encoded before embedding.

(ii) MV values directly obtained from the compressed video are locally optimal, which means data hiding on MVs will shift the local optimal MVs to nonoptimal.

2.2 Add-or-Subtract-One MV Value-Based Steganalysis

The add-or-subtract-one (AoSO) operation on MVs is then presented to analyze the influence produced by Steganography, followed by the extraction of new AoSO feature. Finally, compared to existing features, the universal applicability of AoSO feature is analyzed.

Figure 1 shows a generic structure of the inter-MB coding. The distortion introduced by inter-MB coding is represented by the difference between current MB and reconstructed MB, and is mainly due to quantization and truncation. Since the distribution of the 2D-DCT coefficients of PE can be approximately modeled with the Laplacian probability density function (PDF) as

$$f_y(y) = \frac{\alpha \exp(-\alpha|y|)}{2} \tag{1}$$

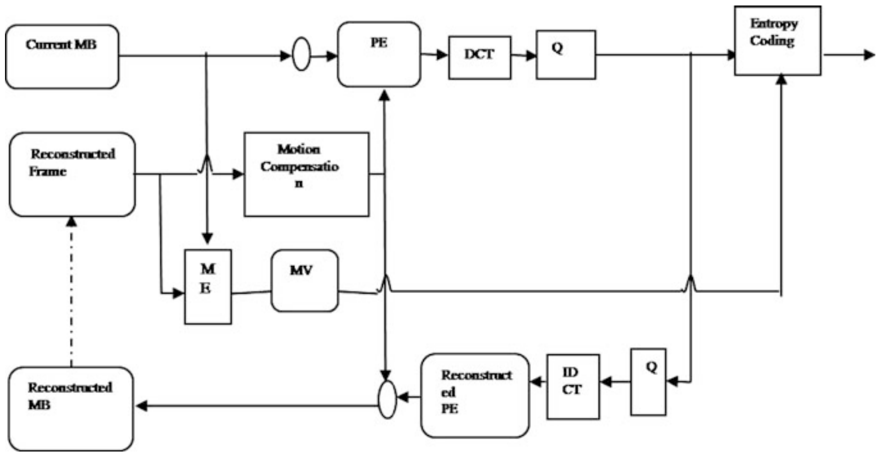


Fig. 1 Inter-MB coding generic structure

where α is the parameter of the distribution, and it is mainly correlated with the accuracy of the motion estimation. The coefficients y are then quantized to get

$$\tilde{y} = iQ; \text{ if } y \in \left[\left(i - \frac{1}{2}\right)Q, \left(i + \frac{1}{2}\right)Q\right] \tag{2}$$

where Q is the quantization step and i is an integer. The probability of the quantized coefficients can be calculated by

$$P_i = \int_{\left(i-\frac{1}{2}\right)}^{\left(i+\frac{1}{2}\right)} f_y(y)\mathrm{d}y \tag{3}$$

Thus the PDF of the residual signal introduced by quantization is

$$f_z(Z) = f_{|y \simeq y|}(|y \simeq y|) \tag{4}$$

and

$$E[Z] = \frac{\tanh\left(\frac{\alpha Q}{4}\right)}{\alpha} \tag{5}$$

$E[\varepsilon] \in (0, Q/4)$ is positively associated with Q, and is negatively correlated. Since the distortion of the DCT coefficients is related to α and simultaneously Q, the distortion of the SAD is also related to α and Q. The larger the Q is, the smaller the α is, and the more serious the distortion of the SAD. The worst case happens when motion estimation is much inaccurate (e.g., the video content is fast moving and of complex texture, the ME method is unreasonable) and DCT coefficients are compressed with a large *Q*uantization step (e.g., the bit rate of the feature is situated too little, or the quantization step is restricted to an area around a huge quality).

3 Results and Discussions

IBP frames chosen for embedding are shown in Fig. 2. Motion vectors estimated are shown in Fig. 3. Macro-block-based processing and frame after embedding data are shown in Figs. 4 and 5. Data embedding process in frame by frame and PSNR between original image and reconstructed image are shown in Figs. 6 and 7.

The PSNR is most normally utilized as quality estimation for lossy packed pictures. The PSNR is the proportion of the maximal force of unique picture and the clamor force of misshaped picture. It is mentioned in the logarithmic area in light of the fact that the forces of signs are normally in a wide element range.

The most noteworthy contrast between the lapse sensibility and the SSIM measurements is the extraction of auxiliary data. The luminance we see in a scene is the result of the light and the reflectance; however, the structure of an item is independent of the brightening. As the objective is to concentrate the basic data from items in a picture, we wish to isolate the impact of the brightening. At the end of the day, the auxiliary data we consider ought to be autonomous of the luminance and the differentiation. In diagram *X* and *Y* are input signals to be measured. First, their luminance is compared. Second, the mean intensities are removed from each signal such that $\sum_{i=1}^{N} x_i = 0$ and $\sum_{i=1}^{N} y_i = 0$, and the signal contrast is estimated

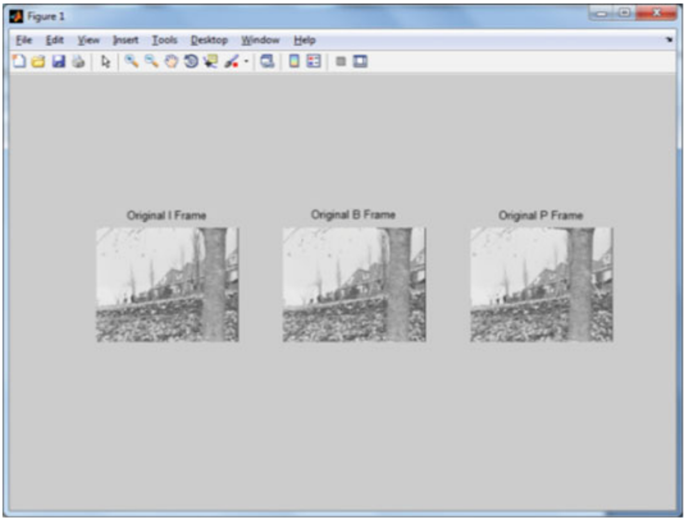


Fig. 2 IBP frames chosen for embedding

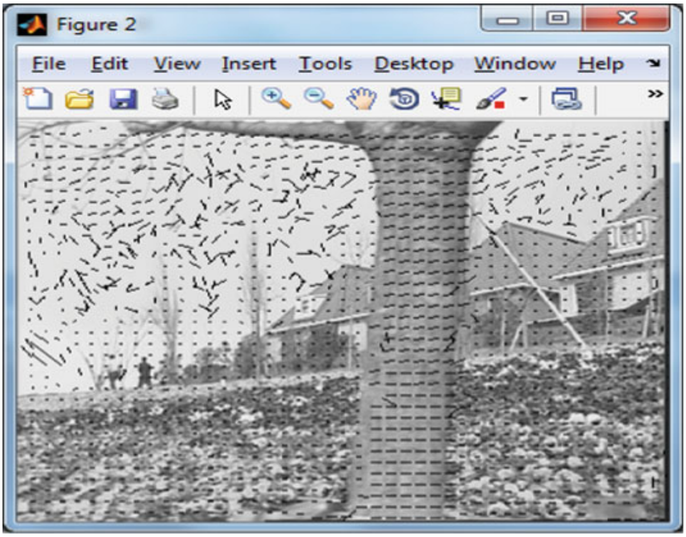


Fig. 3 Motion vectors estimated

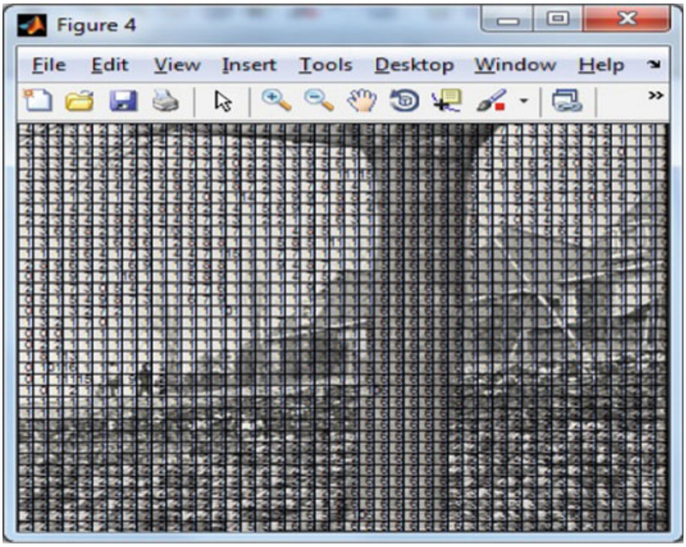


Fig. 4 Macro-block-based processing

by the standard deviations. Third, every sign is standardized by partitioning its standard deviation, so that the two signs being thought about have both unit standard deviations. Next, the structure correlation is led on the two standardized signs. At last, the three components luminance, difference, and structure

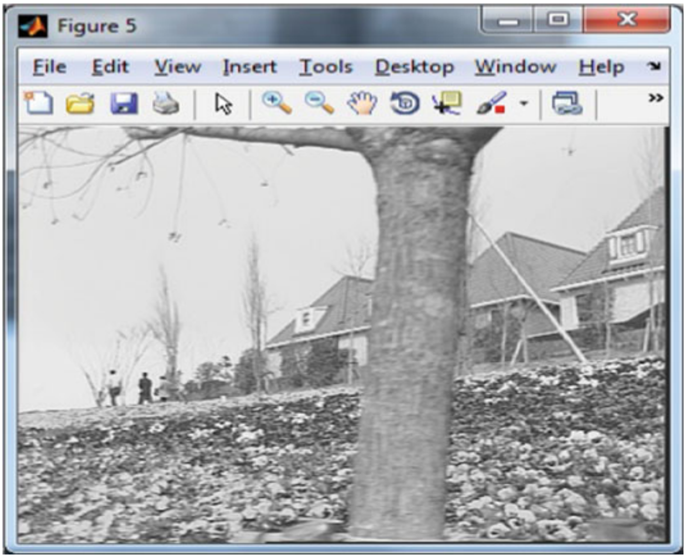


Fig. 5 Frame after embedding data

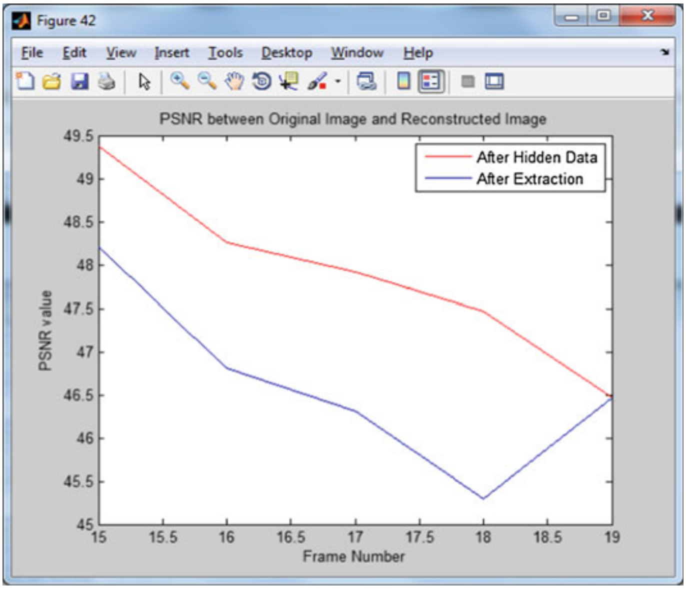
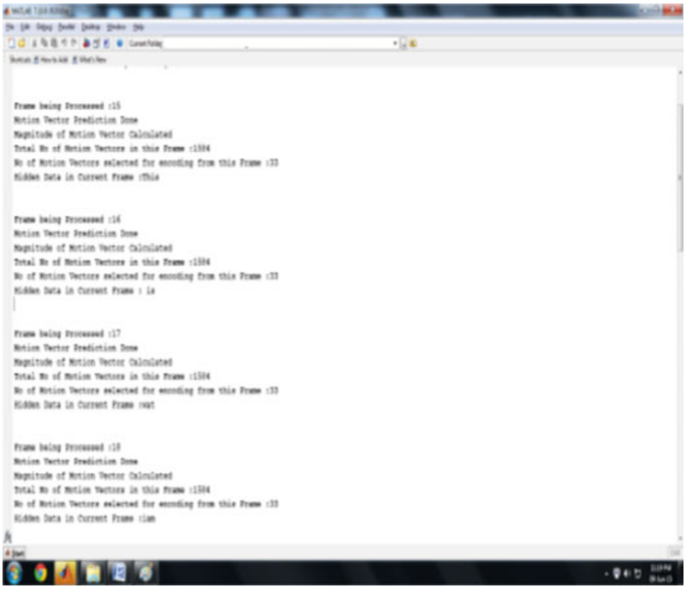


Fig. 6 Data embedding process in frame by frame



```
Frame being Processed :15
Motion Vector Prediction Done
Magnitude of Motion Vector Calculated
Total No of Motion Vectors in this Frame :1584
No of Motion Vectors selected for encoding from this Frame :33
Hidden Data in Current Frame :This


Frame being Processed :16
Motion Vector Prediction Done
Magnitude of Motion Vector Calculated
Total No of Motion Vectors in this Frame :1584
No of Motion Vectors selected for encoding from this Frame :33
Hidden Data in Current Frame : is


Frame being Processed :17
Motion Vector Prediction Done
Magnitude of Motion Vector Calculated
Total No of Motion Vectors in this Frame :1584
No of Motion Vectors selected for encoding from this Frame :33
Hidden Data in Current Frame :not


Frame being Processed :18
Motion Vector Prediction Done
Magnitude of Motion Vector Calculated
Total No of Motion Vectors in this Frame :1584
No of Motion Vectors selected for encoding from this Frame :33
Hidden Data in Current Frame :iam
```

Fig. 7 PSNR between original image and reconstructed image

examination are joined together to yield a general closeness measure. Here, the correlation capacities ought to be characterized such that they can fulfill the three after conditions that ought to go to a picture structure.

A. Luminance comparison
B. Contrast comparison
C. Structure comparison

4 Conclusion and Future Scope

The work presented in this paper makes use of the fact that most ME methods aim at searching at least the locally optimal MV value, as well as the evidence that MV-based steganography has slight influence on the SAD. The AoSO operation on MVs is employed to observe whether the actual MV is locally optimal and how the actual SAD deviates from the locally optimal one. Features based on the AoSO operation are extracted for steganalysis.

The motion value plays a crucial role in analyzing the difference between the rated, which allows us to focus on the difference between the actual SAD and the locally optimal SAD after the adding or subtracting one operation on the motion value. Finally, to perform the classification and extraction process-based motion vectors, two feature sets are been used to complete this process based on the fact that most motion vectors are locally optimal for most video codecs. The proposed method succeeds to meet the practical application requirement and simultaneously succeed in detecting the steganalysis in videos compared to conventional approaches reported in the literature.

References

1. J. Kodovsky, J. Fridrich, and V. Holub, “Ensemble classifiers for steganalysis of digital media,” IEEE Trans. Inf. Forensics Security, vol. 7, no. 2, pp. 432–444, Apr. 2012.
2. J. Fridrich and J. Kodovsky, “Rich models for steganalysis of digital images,” IEEE Trans. Inf. Forensics Security, vol. 7, no. 3, pp. 868–882, Jun. 2012.
3. W. Luo, Y. Wang, and J. Huang, “Security analysis on spatial ± 1 steganography for JPEG decompressed images,” IEEE Signal Process. Lett., vol. 18, no. 1, pp. 39–42, Jan. 2011.
4. F. Jordan, M. Kutter, and T. Ebrahimi, “Proposal of a watermarking technique for hiding data in compressed and decompressed video,” ISO/IEC Document, JTC1/SC29/WG11, Stockholm, Sweden, Tech. Rep. M2281, Jul. 1997.
5. D. Y. Fang and L. W. Chang, “Data hiding for digital video with phase of motion vector,” in Proc. IEEE Int. Symp. Circuits Syst., May 2006, pp. 1422–1425.
6. H. Aly, “Data hiding in motion vectors of compressed video based on their associated prediction error,” IEEE Trans. Inf. Forensics Security, vol. 6, no. 1, pp. 14–18, Mar. 2011.
7. Y. Cao, X. Zhao, D. Feng, and R. Sheng, “Video steganography with perturbed motion estimation,” in Proc. 13th Int. Conf. IH, vol. 6958, 2011, pp. 193–207.

Image Resolution Enhancement Technique Using Lifting Wavelet and Discrete Wavelet Transforms

M. Venkateshwar Rao and V. Bhagya Raju

Abstract This paper presents a technique that used to enhance the image resolution. The input low resolution (LR) image is decomposed into low frequency (LF) and high frequency (HF) sub-bands by using lifting wavelet transform (LWT) and discrete wavelet transform (DWT). The resolution enhancement is done by interpolating HF sub-bands that are generated using LWT from the LR image. The HF sub-bands of DWT are used to preserve the edges of the images. The operation is done on HF sub-bands of LWT and DWT. The HF sub-bands that are generated from the DWT are interpolated and perform addition to the respective interpolated HF sub-bands of LWT. The HF sub-bands that are modified and the LR image will again interpolate and undergoes the inverse lifting wavelet transform (ILWT) to reconstruct the high resolution (HR) image. This implemented method results efficient over conventional methods.

Keywords LWT · Interpolation · DWT and ILWT

1 Introduction

Present generation image resolution enhancement became crucial factor to get maximum details from an image. If the number of pixels per line increases then the resolution also increases. This can be done by using interpolation of an image. A new technique is used to enhance image pixels per line is using wavelet transforms as LWT and DWT. In this method, a grayscale image is used to as an input image to enhance its resolution. Lifting scheme [1, 2] is used for enhancing the

M. Venkateshwar Rao (✉)
DECE, GNITC, Ranga Reddy, Telangana, India
e-mail: venkateshwarrao06@gmail.com

V. Bhagya Raju
JNTUH, Hyderabad, Telangana, India
e-mail: vbhagya01@gmail.com

H.S. Saini et al. (eds.), *Innovations in Computer Science and Engineering*,
Advances in Intelligent Systems and Computing 413,
DOI 10.1007/978-981-10-0419-3_28

image resolution is for perfect reconstruction and efficient computation. Discrete wavelet transform [2, 3] is used at middle stage for preserving edges. The wavelet transform of an image of level one was decomposed into four sub-band images, one of the sub-band image is a low frequency sub-band image and the remaining three image sub-bands are high frequency sub-band images and those high frequency sub-band images contains horizontal, vertical, and diagonal details coefficient.

The idea of regression based approach using bicubic-based interpolation is done using the example-based learning for single-image super-resolution [4]. This method will undergo kernel ridge regression-based technique by factor of two to get enhanced details and smooth edges of an image. This approach is used to enhance the image quality and details at the output. The resultant HR image size is 2γ times greater than LR image. This shows the better results over the conventional methods.

2 Implemented Method

In this implemented technique, the interpolation of the HF sub-band is done using a regression-based approach. The basic image interpolation results un-sharp details. The regression approach uses two methods kernel ridge regression and sparse basis solution. The sparse basis is found by combining kernel matching pursuit and gradient descent. This preserves the major edges and curves (Fig. 1).

The input LR image is decomposed into four sub-bands using LWT [1]. The wavelet transform uses level one of the LWT of LR image. The filter used for decomposition and reconstruction of wavelet transform is 'cdf4.4' type

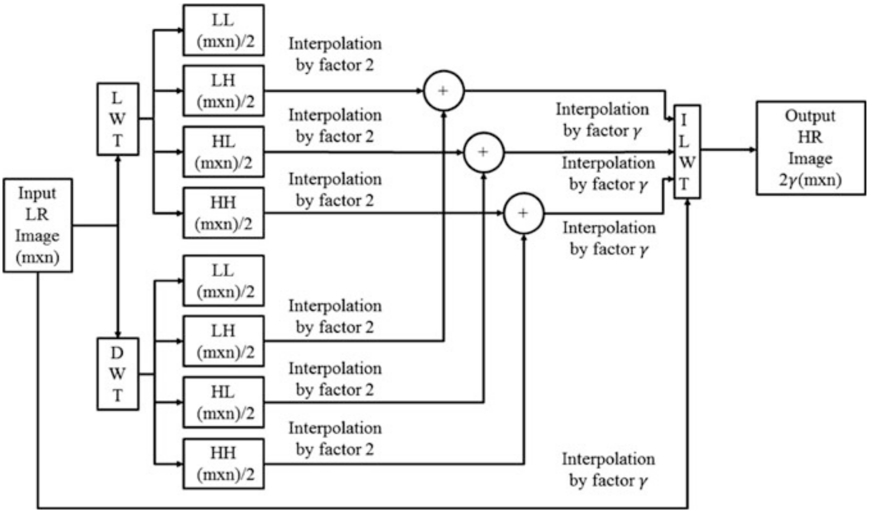

LL (mxn)/2
LH (mxn)/2
HL (mxn)/2
HH (mxn)/2
L W T
Input LR Image (mxn)
Interpolation by factor 2
Interpolation by factor 2
Interpolation by factor 2
+
+
+
Interpolation by factor γ
Interpolation by factor γ
Interpolation by factor γ
I L W T
Output HR Image 2γ(mxn)
D W T
LL (mxn)/2
LH (mxn)/2
HL (mxn)/2
HH (mxn)/2
Interpolation by factor 2
Interpolation by factor 2
Interpolation by factor 2
Interpolation by factor γ

Fig. 1 Implemented algorithm

(Cohen–Daubechies–Feauveau 4.4 as lift wave). For the level one of the DWT of LR image decomposition wavelet filter used as ‘bior4.4’ (Bi-orthogonal 4.4wavelet filter).

The HF sub-bands that are decomposed form LWT and DWT are half in size when compared with input LR image. The kernel ridge regression-based interpolation is done parallel to HF sub-bands of LWT and DWT. Now these HF component sub-bands are equal in size with input LR image. The interpolated HF sub-bands that are form LWT and DWT were added with respective sub-bands. The modified and adjusted HF sub-bands are again interpolated by a factor of 2. The input LR image is interpolated by a factor of 2. The reconstruction of HR image is obtained by applying ILWT on interpolated LR image and resultant HF sub-bands. The HR image is now 4 times greater than the input LR image.

3 Results

The implemented method is applied on well-known images like Lena, Baboon, and Peppers. These images are taken from USC-SPI image database [5]. All programs in this experiment are written in MATLAB (R2014a). The resolution factor was 4 times of the input image. The resolution enhanced form the images are from 128 × 128 to 512 × 512. For comparison of existing methods PSNR metric is used.

To calculate PSNR, the mean square error (MSE) of image must be calculated and this can formulated as in Eq. (1). The PSNR of an image is formulated as in Eq. (2).

$$\mathrm{MSE} = \frac{1}{\mathrm{MN}} \sum_{m=0}^{M-1} \sum_{n-0}^{N-1} [\mathrm{Image}(m,n) - \mathrm{Reference}(m,n)]^2 \tag{1}$$

$$\mathrm{PSNR}\,(\mathrm{dB}) = 10 \times \log_{10}\left(\frac{255 \times 255}{\text{Mean Square Error}}\right) \tag{2}$$

The PSNR comparison of the enhanced image form 128 × 128 to 512 × 512 by this method shows the best results over existing methods are tabulated in Table 1.

The outputs of implemented method images of each of size which input image size of 128 × 128 is enhanced the image resolution to the image of size of 512 × 512. The resolution enhanced image of Lean, Baboon and Peppers images are shown in Fig. 2.

The PSNR comparisons of enhanced image form 256 × 256 to 1024 × 1024 is shows in Table 2. The implemented method images of Wall and Pentagon are shown in Fig. 2.

Table 1 PSNR (dB) results of images enhanced from 128 × 128 to 512 × 512 of implemented method comparing with existing methods

Method/images	PSNR (dB)			
	Lena	Baboon	Peppers	Elaine
Bilinear	26.34	20.51	25.16	25.38
Bicubic	20.86	20.61	25.66	28.93
NEDI [6]	28.81	21.18	28.52	29.97
HMM [7]	28.86	21.47	29.58	30.51
DWT SR	34.79	23.29	32.19	32.73
DWT and SWT SR [3]	34.82	23.87	33.06	35.01
LWT and SWT [8]	34.91	28.92	33.06	34.95
LWT and DWT implemented method	39.01	38.14	38.78	40.37

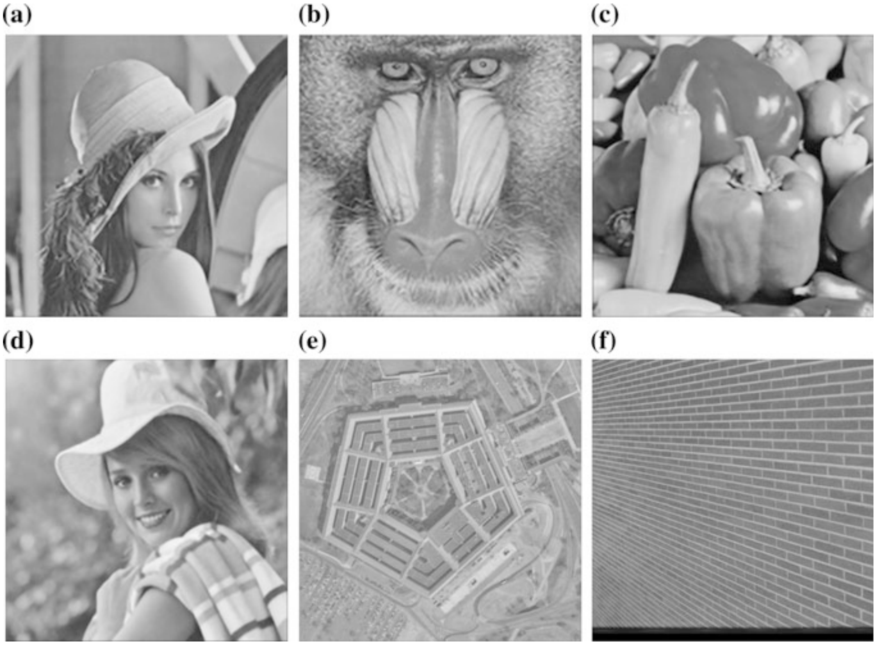


Fig. 2 Output images of implemented method. **a** Lena (512 × 512). **b** Baboon (512 × 512). **c** Peppers (512 × 512). **d** Elaine (512 × 512). **e** Pentagon (1024 × 1024). **f** Wall (1024 × 1024)

Table 2 PSNR (dB) results of images enhanced from 128 × 128 to 512 × 512 of implemented method comparing with existing methods

Methods/images	PSNR (dB)	
	Wall	Pentagon
Bilinear	29.27	28.61
Bicubic	30.24	28.99
DWT SR	32.45	29.67
DWT and SWT SR	33.18	31.05
LWT and SWT	36.98	32.01
LWT and DWT implemented method	37.95	40.02

4 Conclusion

The implemented method on different images to enhance the image resolution by modifying the HF sub-bands of LR image. The interpolation of high frequency sub-bands of the LR-input image those sub-bands were decomposed from the LR image using LWT and DWT methods. The interpolated high frequency sub-bands from LWT are modified and adjusted by summing the interpolated high frequency sub-bands from DWT. These modified HF sub-bands are again interpolated parallel and interpolated LR-input image undergoes reconstruction of image of size which is four times greater than the input LR image by using the ILWT to get high resolution of image. This method experimented on different images, such as Lena, Baboon, Peppers, Elaine, and Wall and Pentagon. Finally, this implemented method shows the best results when compared with existing methods.

References

1. Sweldens W, "Wavelets and the Lifting Scheme: A 5 Minute Tour," Int. Journal ZAMM Zeitschrift fur Angewandte Mathematik und Mechanik, vol. 76 no. 2, pp. 41–44, 1996.
2. The Math Works, Inc. Database resource [Online] Available: http://www.mathworks.in/help.
3. Hasan Demirel and Gholamreza Anbarjafari, "Image Resolution Enhancement by Using Discrete and Stationary Wavelet Decomposition", IEEE Trans. Image Procss., vol. 20, no. 5, 1458–1460, May 2011.
4. K. I. Kim and Y. Kwon, "Example-based learning for single-image super-resolution", in *Proc. DAGM*, pp. 456–465, 2008
5. A. Weber, USC-SIPI Image Database resource [Online] Available: http://www.sipi.usc.edu/database/database.php.
6. X. Li and M. T. Orchard, "New edge-directed interpolation," IEEE Trans. Image Process., vol. 10, no. 10, pp. 1521–1527, Oct. 2001.
7. K. Kinebuchi, D. D. Muresan, and R. G. Baraniuk, "Wavelet based statistical signal processing using hidden Markov models," in Proc. Int. Conf. Acoust., Speech, Signal Process., 2001, vol. 3, pp. 7–11.
8. Agrawal, M; Dash, R, "Image Resolution Enhancement by using Lifting and Stationary Wavelet Transform" IEEE Trans. Image Procss, pp. 322–325, 2014

A Real-Time Implementation of Face and Eye Tracking on OMAP Processor

Vijayalaxmi Biradar and D. Elizabath Rani

Abstract The real-time implementation of embedded image processing applications needs a fast processor. Eye recognition is an important part of image processing systems such as driver fatigue detection system and eye gaze detection system. In these systems, a fast and accurate real-time implementation of face and eye tracking is required. Hence, a new approach to determine and track face and eye on live images is proposed in this paper. This proposed method is implemented and successfully tested in laboratory for various real-time images with and without glasses captured through Logitech USB Camera of 1600 × 1200 pixels @ 30 fps. The method is developed on 1 GHz open multimedia applications platform (OMAP) processor and the algorithm is developed using OpenCV libraries. The success rate of the proposed algorithm shows that the hardware has sufficient speed and accuracy, which can be used in real time.

Keywords DM3730 · Eye · Face · Logitech · OpenCV

1 Introduction

Wierwille et al. [1] proposed monitoring changes in physiological characteristics like ECG, EEG, skin temperature, and head movement to estimate driver fatigue. The drawback of this approach is that it causes distraction, nonrealistic, and disturbance to the driver. Artaud et al. [2] and Mabbott et al. [2] proposed a method of sensing driver response by placing sensors on steering wheel and back of the seat. The drawback of this approach is that it fails if driver wears gloves and performance of sensors placed on back seat reduces with time. Boyraz et al. [3] proposed a

V. Biradar (✉)
Vignan Institute of Technology and Science, Hyderabad, Deshmukhi, India
e-mail: laxmi81181@gmail.com

D. Elizabath Rani
Gitam Institute of Technology, GITAM University, Visakhapatnam, India

H.S. Saini et al. (eds.), *Innovations in Computer Science and Engineering*,
Advances in Intelligent Systems and Computing 413,
DOI 10.1007/978-981-10-0419-3_29

method of sensing vehicle response to measure uncertainty in steering wheel. The drawback of this approach is that it is limited to vehicle type and driver condition. Mabbottt et al. [2] proposed a method where driver response is monitored requesting the driver to send feedback continuously. The drawback of this approach is that driver gets tiresome and feels annoying.

All the above-discussed approaches are intrusive system of driver fatigue detection system. Few limitations of intrusive methods are as follows: system is complex, cannot be placed easily, causes disturbance, poor performance, non-reliable, and produces noise. The solution is nonintrusive system. The basic approach in nonintrusive system is analysis of face. The first symptom of fatigue appears in eye and than in mouth. Lot of research has been done to analyze the facial features to estimate driver fatigue based on eye blink rate and yawning.

Parmar [4] proposed driver drowsiness detection system based on eye blink rate, the success rate is 80 %. The drawbacks of the system are poor illumination, unable to track eyes of person wearing spectacles, and fails if one or both eyes are closed. Gallagher [5] proposed driver alert system for road safety, the problem with the system is that it takes 8 s to process each frame which is not desirable in real-time implementation.

Achieving high performance is a challenging task in the field of image processing. Many useful image processing algorithms are described quite compact with few operations, and these operations need to be repeated over large amount of data and usually demand special requirements. Meeting all these requirements is a challenging task [6]. Powerful digital signal processors are needed to handle the image processing complexity but these are expensive, inflexible, and consume more time for development. To meet the complexities of image processing and computer vision, high-performance computing (HPC) is most commonly used technology. But HPC has failed to satisfy cost, size, or power consumption requirements which are key issues in image processing applications [6, 7]. Multiply–accumulate (MAC) operations can be performed easily on DSP. Wisdon and Lee, Koo et al., in 2007 have proposed implementation of image processing algorithms on FPGA. So in order to carry out any image processing-based applications it is required to analyze the performance of hardware. In this paper, a new approach for carrying out image processing applications on open-source platform with high speed and low power consumption on OMAP processor is discussed.

The rest of the paper is organized as follows: Sect. 2 discusses about OMAP processor, Sect. 3 explains the steps required to port the operating system, Sect. 4 discusses algorithm to track face and eye, Sect. 5 discusses about result, and conclusions are drawn in Sect. 6.

2 OMAP Processor

Open multimedia applications platform (OMAP) is a series of image/video processors developed by Texas Instruments. They are a category of proprietary system on chips (SoCs) for portable and mobile multimedia applications. The OMAP family is categorized into three product groups based on performance and intended application. High-performance application processors are intended to use in smartphones. Basic multimedia application processors are used in low-cost consumer products. Integrated modem and application processors are used in low-cost cell phones [8].

The DM3730 generation of high-performance, digital media processors is based on the enhanced media device architecture and is integrated on Texas advanced hardware. This architecture is designed to provide best-in-class ARM Cortex A8 and graphics performance while delivering low power consumption. The BeagleBoard is low-power open-source hardware, single-board computer produced by Texas Instruments. The BeagleBoard is also designed with open-source software development in mind, and as a way of demonstrating the Texas Instrument's OMAP3730 system-on-a-chip [8]. The block diagram of BeagleBoard DM3730 processor is shown in Fig. 1.

The BeagleBoard can be powered using either USB OTG port of 5 V DC supply. The architecture is designed to provide best-in-class video, image, and graphics processing sufficient to various applications. The processor supports high-level operating systems such as Windows CE, Linux, Symbian, and others. The BeagleBoard-xM has a faster CPU core (clocked at 1 GHz), RAM (512 MB), onboard Ethernet jack, and 4-port USB hub. The BeagleBoard-xM lacks the onboard NAND and therefore requires the operating system and other data to be stored on a microSD card [9].

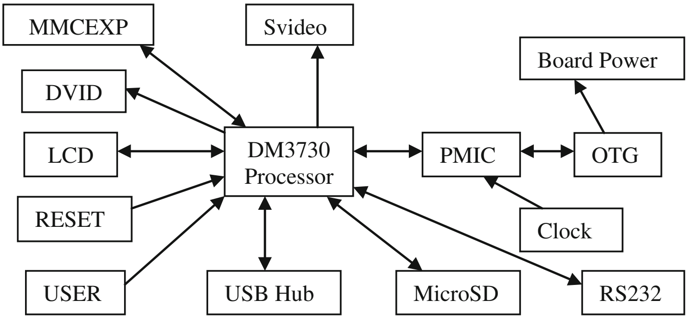


Fig. 1 Block diagram of DM3730 processor

3 Steps to Install Operating System

The operating system installed on microSD card is Ubuntu 11.10. Following are the steps used to install operating system on microSD card. The preferred microSD card is SanDisk to install operating system. The software required to perform image processing algorithm to track face and eye are Python and OpenCV.

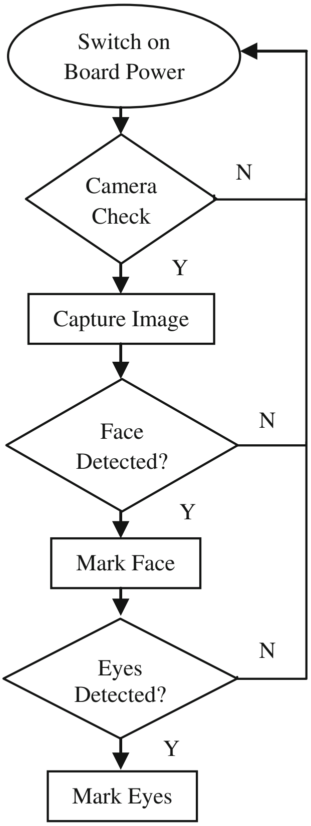
Step(i):	Pre-requirements—First insert the microSD in Linux System, and before installing OS make sure that the required softwares are installed onto Linux System.
Step(ii):	Identify the location of microSD card.
Step(iii):	Download the stable release of Ubuntu [10].
Step(iv):	Checksum is used to check whether the OS is downloaded properly or not.
Step(v):	The downloaded OS is a tar file, similar to Zip file in windows.
Step(vi):	Last step is to write the SD card which installs the OS onto the card.
Step(vii):	After installing the OS, remove the SD card from Linux system, insert it in BeagleBoard and power up the board with 5 V DC supply.
Step(viii):	Internet connection is required to upgrade and update the OS.
Step(ix):	After step(viii), the graphical user interface (GUI) is to be installed, in order to use this board as a single-board computer.

After performing all the nine steps successfully, the USB camera is connected verify. Since, Ubuntu is a Linux version and it is open source; hence, all the USB drivers will come as a part of OS. In order to check the working of camera, light weight package Luvcview is installed and verified the working of camera. The steps to install python and OpenCV [11] are successfully installed and tested.

4 Algorithm for Face and Eye Tracking

The image processing algorithm for tracking face and eye from the captured images through Logitech USB camera is shown in Fig. 2. Images are captured through USB camera @30 fps. The video is captured using VideoCapture which uses Video4Linux to capture raw stream from the USB camera. This raw stream is then wrapped with H264 header using FFMPEG. V4L or Video4Linux is an application programming interface for video capture which supports many USB cameras on Linux operating system. OMAP is an open-source hardware which supports all OpenCV libraries. These libraries are useful to perform image processing-based applications such as human–computer interface, biometrics, etc. [12]. With this VideoCapture, it streams frames @30 fps of 640 × 480 sizes with 17 % of CPU utilization. The captured images are processed to track face and eye. Some existing methods to track face and eyes are skin segmentation [13], template matching [14], and neural network approach [15].

Fig. 2 Flow chart for face and eye tracking



In this algorithm, face and eye classifiers are trained with two sets of images. One set of images include non-object data such as background, chair, and scene which does not have facial features. This set of images is considered as negative images. The other set of images consists of faces and eyes with different orientations, illuminations, sizes, and different age groups. This set of images which contain one or more instances of the object is called as positive images. Around 100 negative images are used from VITS database and 100 positive images from GTAV Database [9] and VITS database. The GTAV database consists of 44 images of 27 different persons taken at different angles of 0°, 30°, 45°, 60°, and 90° from a frontal view. The resolution of these images is 240 × 320 and they are in BMP format. The VITS database consists of images captured under different illumination conditions of 11 different persons. The resolution of these images is 640 × 480 and they are in JPEG format. Two separate classifiers are trained, one for the face and one for the eyes. These classifiers are loaded to detect face and eyes on the video captured from the Logitech USB camera.

5 Results

The real-time face and eye tracking algorithms developed using Python and OpenCV libraries are implemented on DM3730 processor that is successfully tested on various images with and without glasses in laboratory. The total of 100 images each with and without glasses is tested and results are tabulated in Table 1. The original images captured through Logitech USB camera under different illumination conditions are shown in Fig. 3. Face and eye detected with and without glasses are shown in Fig. 4. Few failure cases are shown in Fig. 5. The algorithm fails to detect eyes due to reflection of glasses.

Table 1 Results

Type of images	Total images used for testing	Face detected success rate in %	Eye detected success rate in %
With glasses	100	100	96
Without glasses	100	100	100

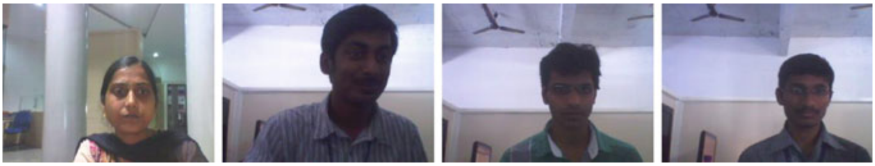


Fig. 3 Original images captured through logitech USB camera

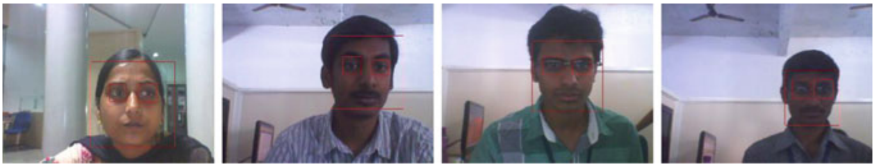


Fig. 4 Face and eye detected images

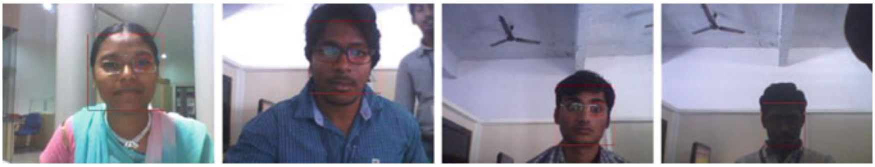


Fig. 5 Failure images

The proposed algorithm is able to detect face and eyes under different illuminations. This proposed algorithm shows better results on different images captured under different illumination conditions, and also works on images with glasses when compared to existing methods, and the success rate is tabulated in Table 1.

The tabulated results show that face detection with and without glasses works perfectly without any constraints. Eye detection without glasses works perfectly but images with glasses have some constraints like lighting effects but algorithm has detected one eye successfully.

6 Conclusion

Face and eye tracking is implemented on DM3730 processor. The purpose of this paper is to make use of resources available and inspire to work on open-source platform. The DM3730 processor architecture has best in class and has CPU with 1 GHz speed which is generally the major requirement to perform image processing applications. The proposed algorithm is successfully tested on various images with and without glasses. The method works with reasonable lighting conditions. Further, the main aim of this research is to develop a system which can detect driver's fatigue based on eye blink rate. The authors continue to work on open-source platform and contribute to the research in the field of image processing. The success rate of proposed algorithm is 100 % for images without glasses and 96 % with glasses.

References

1. W. W. Wierwille, S. S. Wreggit, C. L. Kirn, L. A. Ellsworth, and R. J. Fairbanks III, "Research on vehicle-based driver status/performance monitoring: development, validation, and refinement of algorithms for detection of driver drowsiness," National Highway Traffic Safety Administration, U.S. DOT Tech Report No. DOT HS 808 247, (1994).
2. Artaud et al., Mabbott et al., Lavergne et al., Vitabile et al., Eskandarian. A & R. Sayed in 2005, "Monitoring the response of drivers", (1994), (1999), (1996), (2008), (2005).
3. Boyraz. P., Leicester, Hansen J.H.L, Sensing of Vehicle response, (2008).
4. Neeta Parmar, Drowsy Driver Detection System, in (2002).
5. Martin Gallagher, "Development of a driver alert system for road safety", in (2006).
6. Almudena Lindoso and Luis Entrena, Hardware Architectures for Image Processing Acceleration, Image Processing, Yung-Sheng Chen (Ed.), ISBN:978-953-307-026-1, InTech, doi:10.5772/7066, (2009).
7. Beymer D J, "Face Recognition under varying pose", IEEE Proceedings of International Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, Washington, USA, pp. 756–761, (1994).
8. Coley, Gerald, "Take advantage of open-source hardware", EDN, (2011).
9. Paul, Ryan (2008-08-01), "TI launches hackable Beagle Board for hobbyist projects", 16 January (2010).
10. http://rcn-ee.net/deb/rootfs/oneiric/ubuntu-11.10-r10-minimal-armel.tar.xz.

11. http://tothinkornottothink.com/post/59305587476/raspberry-pi-simplecv-opencv-raspicam-csi-camera.
12. Adolf F., How to build a cascade of boosted classifiers based on Haar like features OpenCVs Rapid Object Detection (2003)
13. Sudhakar Rao P, Vijayalaxmi, Sreehari S, “A new procedure for segmenting eyes for human face”, IJ-ETA-ETS, Volume 4, Issue 2, ISSN: 0974-3588, pp. 210–213, July-Dec (2011).
14. Vijayalaxmi, Sreehari, “ Knowledge based template for Eye detection”, National Conference on Microwave Antenna and Signal Processing, pp. 90, April (2011).
15. Vijayalaxmi, Sudhakara Rao P, Sreehari S, “Neural Network Approach for eye detection”, The Second International Conference on Computer Science, Engineering and Applications (CCSEA), Proceedings Volume Editors: David C. Wyld, Jan Zizka, Dhinaharan Nagamalai ISBN:978-1-921987-03-8, pp. 269–281, May (2012).

A Study on Cyber Security, Its Issues and Cyber Crime Rates in India

Saurabh Mishra, Saru Dhir and Madhurima Hooda

Abstract In the current technological era, use of computers becomes an essential part of our lives. But this part is also affected by a new breed of security known as cyber security. It is a global issue that arises by different organisations. This paper presents the global cyber security scenario, cyber security and its practices, and firms who are major stakeholders in the cyber security. At the end counter measures of cyber crimes, its average rate is calculated in India during the years 2009–2013.

Keywords Cyber security · Issues · Cyber cases · Cyber crime

1 Introduction

Today, cyber security has become a global issue attracting widespread concern from all across the world from various organisations such as Governments and International Bodies. A simple reason for it: browsing on a trusted website can completely compromise your computer, allowing a hacker to read your sensitive files or worse, and delete them. The term 'hacker' has become a part of our everyday scenario, and projected in nearly daily headlines.

Global cyber crime incidents are increasing at a rapid rate which is not only unprecedented but also alarming, for much of today's critical infrastructure like electricity, water, gas and secure data, like banking is completely computer based. In such a scenario, a cyber attack on the stock market would probably affect more people than a bomb in a marketplace. Governments are hiring computer security

S. Mishra (✉) · S. Dhir · M. Hooda
Amity University Uttar Pradesh, Noida, India
e-mail: mishrasaurabh95@gmail.com

S. Dhir
e-mail: sdhir@amity.edu

M. Hooda
e-mail: mhooda@amity.edu

H.S. Saini et al. (eds.), *Innovations in Computer Science and Engineering*,
Advances in Intelligent Systems and Computing 413,
DOI 10.1007/978-981-10-0419-3_30

professionals to counter the growing menace. Every day, new advancements in cyber security are made. Ethical hackers are a very important part of the cyber security movement throughout the global scenario. Cyber security is being made available to all computer users. Cyber security is thus an integral part of the IT scenario, today and overlooking it can gravely undermine the IT sector itself. Cyber security and IT sector are closely entwined where the cyber security is for everything, from computer viruses, Trojans, worms and other malicious programmes; to malicious characters like hackers who intend to get into your computer system; and to security protocols, policies and compliance and regulatory concerns. Here, an effort has been made to give a research-oriented analysis and point of view cyber security, ethical hacking and why we need both to keep the IT sector going smoothly. An effort has been made to do an analysis of cyber security issues, firms that are providing cyber security services, and measures for effective cyber security in India.

2 Literature Review

Cyber security today has become a global issue, for it has now become a matter of economic importance, privacy in society and national security where the cyber network has become a part of the national economy. Security models suggest that to counter this in an effective manner and keep the attacks in check, it is recommended to place intrusion detection devices at the weak points of the networks [1].

Security system has its disadvantages as well; new techniques are being developed today like applications that take part in their own defence [2]. An effective counter measure for protection is the use of DNS health indicators where a proper method and procedure defines the way for measuring the DNS security levels. Such a step can help secure the critical infrastructure during a cyber attack or cyber war scenario [3].

As an analysis, it was found that undetected malware attack cut off the Royal Air Force and Royal Navy from defence network access, and most of hospitals also had lost their network connectivity due to the same malware attack [4]. Furthermore, research and development has been done to obtain self-reliance; and compliance and enforcement have been applied. A good international cooperation from international organisations like UN has also helped the country in obtaining its objective [5].

Major stakeholders in the computer industry are also the top cyber security firms today such as Hewlett Packard Company, Dell Inc., IBM and Intel Corporation, which are all providing cyber security services to a wide population, whereas Kaspersky Labs and Symantec Corporation are providing antivirus solutions to the general everyday users and providing cyber security solutions. All these firms have been providing cyber security solutions and have been the front runners in the cyber security movement [6]. Meanwhile, people ask today, “Can cyber terrorism really affect our computer systems and jeopardise a Nation’s security?” [7]. Well, the answer to this is “Unfortunately, yes”; probably the reason why all agencies, from FBI and CIA to the Indian RAW, have their own specialised cyber security cells. The US suffered from the largest power outage in history on 14 August 2003, which left one-fifth of the population without power, i.e. about 50 million people for over 12 h [8].

3 Issues Relating to Cyber Security

Cyber security today has several issues: Phishing, pharming and e-mail spoofing. These are all issues faced by people while simply browsing the internet or using services like net banking [9].

One of the biggest and most significant cyber security issues is the one being faced by each nation in the world today: An attack on their critical infrastructure and information grid. The U.S. has setup a Defense Critical Infrastructure Programme (DCIP), where each critical infrastructure is assigned a separate lead agent to provide security [10]. Cyber criminals use address and logos resembling those of trusted organisations like banks to obtain the user’s privacy information like passwords and credit card numbers. A case of a very good cyber security and protection step is that of the Malaysian Government to counter the cyber security issues faced by the country. It was aimed at protecting assets vital to the nation like it is image, defence and security and different sectors through affective governance, a good legislature and regulatory framework, a strong cyber security technology framework and developing a culture of security and capacity building in its citizens [11].

A very stunning and alarming cyber security issue, whose significance is understood by very few, is how easily hacking tools are available today on the internet and can be used to create severe cyber security problems that can cripple even an entire nation. An example for this is the “Eligible Receiver”, exercise conducted by the National Security Agency (NSA) in 1997, where the NSA conducted an experiment by briefing 35 computer hackers to hack and disrupt U.S. National Security Systems and Databases using only software and hacking tools available freely for download on the Internet. The results were appalling, where the 35-man team was able to compromise several security sectors in the U.S. Pentagon and other Government organs [12]. And sadly, this issue can never be kept entirely in check, as such tools for hacking can never be entirely removed from the internet.

4 Survey Results on Cyber Security in India

To give a brief idea of the cyber crime scenario, we will take the case of India, where a survey of cyber crimes committed during the period 2009–2013 was done by the National Crime Records Bureau (NCRB) [13]. The graph as shown in Fig. 1

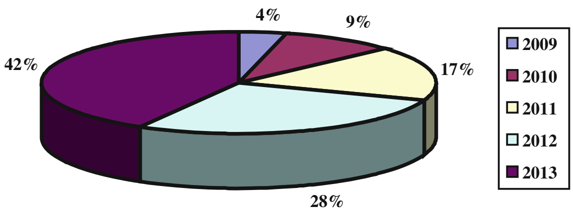


Fig. 1 Cyber crime cases registered under year 2009–2013

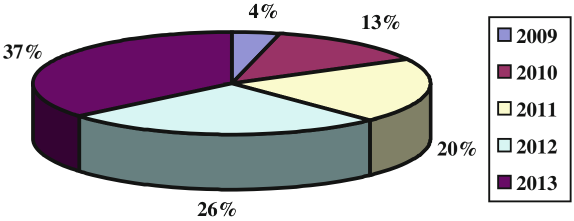
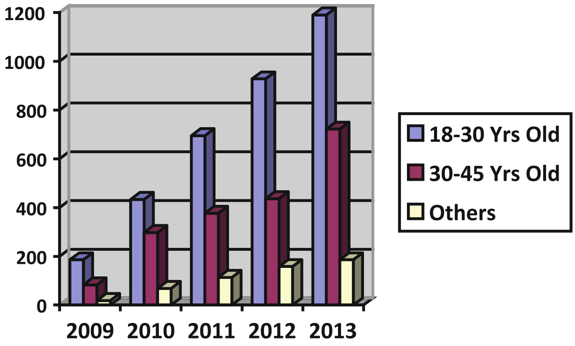


Fig. 2 Persons arrested for cyber crimes under year 2009–2013

Fig. 3 Age-wise breakdown of people arrested for cyber crimes (2009–2013)



represents that how there has been a steady rise in the number of cyber crimes cases registered in India in the past 5 years. From a mere 420 cases in the year 2009, it has risen to 4356 cases in the year 2013, which is more than 10 times the number of cases registered in 2009.

Figure 2 represents how there has been a sharp rise in the number of cyber criminals in the recent years, where, of the total people arrested in the last 5 years, 2098 people were arrested in the year 2013 alone. This is 37 % of the total criminals arrested in the period 2009–2013 for committing cyber crimes.

The graph in Fig. 3 shows how there has been a steady rise in the number of people arrested from all age groups. But one can clearly notice that the maximum number of people arrested for committing cyber crimes is in the age group of 18–30 years, followed by people of the age group 30–45 years.

5 Conclusion

With this paper, we have discussed the global cyber security scenario, where today the critical Infrastructure, people's privacy and internet protection are all under threat from the increasing number of cyber crimes. We have discussed cyber security and its practices, and firms who are major stakeholders in the cyber security scenario. We have also taken up cyber security issues and discussed their countermeasures. A general idea about the number of cyber crimes and criminals arrested in India has also been given through survey results for the years 2009–2013.

References

1. Jessie J Walker, Travis Jones and Roy Blount. Visualization, Modeling and Predictive Analysis of cyber security attacks against cyber infrastructure oriented systems. 978-1-4577-1376-7/11. page no. 82. IEEE (2011).
2. Rick A. Jones, Barry Horowitz. System-Aware Cyber Security. 978-0-7695-4367-3/11. page no. 914. IEEE (2011).
3. Andrea Rigoni, Igor Nai Fovino, Salvatore Di Blasi, Emiliano Casalicchio. Worldwide Security and Resiliency of Cyber Infrastructures: The Role of the Domain Name System. 978-0-615-51608-0/11. page no. 2 (2011).
4. Desire Athow, Trojan Malware Penetrates British Navy Defences–http://www.itproportal.com/2009/01/16/trojan-malware-penetrates-british-navy-defences/ (2009).
5. UN-backed anti-cyber-threat coalition launches headquarters in Malaysia- http://portal.unesco.org/ci/en/ev.php-URL_ID=28464&URL_DO=DO_TOPIC&URL_SECTION=201.html/ (2009).
6. Top 20 Cyber Security Companies 2014- http://www.reportlinker.com/p02148719-summary/Top-20-Cyber-Security-Companies.html (2014).
7. A Congressional Guide: Seven Steps to U.S. Security, Prosperity, and Freedom in Cyberspace - http://www.heritage.org/research/reports/2013/04/a-congressional-guide-seven-steps-to-us-security-prosperity-and-freedom-in-cyberspace. (2013).
8. Major power outage hits New York, other large cities-http://edition.cnn.com/2003/US/08/14/power.outage/ (2003).
9. Alex Roney Mathew, Aayad Al Hajj and Khalil Al Ruqeishi. Cyber Crimes: Threats and Protection. 978-1-4244-7578-0. page no. 16-17. IEEE (2010).
10. Critical Infrastructure Threats and Terrorism. DCSINT Handbook No. 1.02. page no. 4 (2006).
11. Mohd Shamir B Hashim. Malaysia's National Cyber Security Policy: The Country's Cyber Defence Initiatives. 978-0-615-51608-0/11. page no. 2–7 (2011).
12. Gabriel Weimann. Cyberterrorism: The Sum of All Fears?. Studies in Conflict & Terrorism, 28:129–149. Taylor & Francis Inc., page no. 138. (2005).
13. National Crime Records Bureau. Crimes in India (Compendium). http://ncrb.nic.in/ciiprevious/main.htm.

A Concept-Based Model for Query Management in Service Desks

G. Veena, Aparna Shaji Peter, K. Anitha Rajkumari and Neeraj Ramanan

Abstract Thousands of email queries are often received by help desks of large organizations nowadays. It is a cumbersome and time-consuming task to manage these emails manually. Also, the support staff who initially answers the query may not always be technically sound to do this themselves. In that case, they forward the queries to higher authorities, unnecessarily wasting their precious time. A large amount of time and human effort is being wasted for this manual classification and query management process. So, in this paper, we propose a new concept-based semantic classification technique to automatically classify the help desk queries into multiple categories. Our system also proposes an approach for retrieving powerful information related to the queries. In our work, the dataset is represented using a graph model and the concept of ontology is used for representing semantics of data.

Keywords Concept-based · Ontology · OWL · RDF · Graph model · Semantics

1 Introduction

Service desk, also known as help desk or support center provides technical advices, information, and troubleshooting guidance to users. These service desks are the primary access points of users to the concerned organization. Effective management

G. Veena (✉) · A.S. Peter · K.A. Rajkumari · N. Ramanan
Department of Computer Science and Applications,
Amrita Vishwa Vidyapeetham, Amritapuri, India
e-mail: veenag@am.amrita.edu

A.S. Peter
e-mail: aparnashajipeter@gmail.com

K.A. Rajkumari
e-mail: anitha6.rk@gmail.com

N. Ramanan
e-mail: neerajharipad@gmail.com

H.S. Saini et al. (eds.), *Innovations in Computer Science and Engineering*,
Advances in Intelligent Systems and Computing 413,
DOI 10.1007/978-981-10-0419-3_31

of user queries at service desks leads to high user satisfaction, which helps in minimizing the churn rate. Incompetent service desk management leads to pointless wastage of time and resources. That is why constructive query management in service desks becomes a major concern. Most of the service desk management systems follow a query response model. The query, in most cases is in the form of email text. User sends a query, the concerned staffs at service desk reads it and he may forward it to the person handling that particular type of queries. In most of the service desks, this classification is done manually. Manual classification has several problems.

The staff at the front desk may not always be a highly technical person. He may not understand the query completely. In this case, he forwards the query to the next higher official for classification. So it becomes his responsibility to categorize the query, which is actually not a part of his job. If the support person is not experienced enough, there is a chance that he may classify the mail to an incorrect category. Then the query needs to be again forwarded to the concerned staff. This also causes pointless delay in the whole process. So, if we could automate this classification process, this can be avoided.

In this paper, we propose a system which avoids these issues. Also, as the email texts are usually very short, the semantics has to be considered for the classification to be accurate. Conventional keyword-based text classification methods do not consider the semantic dependency between attributes. But the real-world data often involves complex relationships among attributes. To represent text without losing semantics, the traditional text representation techniques like vector space model [1] will not be adequate. So, we use a concept-based graph model for representing text data which involves triplet representation [2], considering the semantics of data as well. To describe in graph model, we use Resource Description Framework (RDF) and to define the semantics of data described using RDF, we use Web Ontology Language (OWL).

The organization of the rest of this document is as follows. Section 2 describes the related works; Sect. 3 describes the proposed solution approach, followed by the experimental results in Sect. 4 and conclusion in Sect. 5.

2 Related Works

Text classification is a research area, which has been undergoing many changes from the past few years. Many research works are being done in this area.

Paper [3] discusses in detail about the challenges in managing help desks and the application of knowledge management techniques in help desk management. They also use an ontology-based approach. The advantage of our system is that we use a concept-based, semantic technique for service desk query management. The disadvantages of Naïve Bayesian approach have been discussed in many researches. The paper in Ref. [4] discusses in detail about Naïve Bayes classifier. Particularly, the paper demonstrates that Naive Naïve works best when the features are completely independent. In our case, Naïve Bayes will not be a better option because it does not consider the semantic relationship among features. Vector space model is an algebraic model, which was the most widely used model for representing text data, as in [1]. Here, each text document is represented as a vector. The presence or absence of a feature, or even the term weights can be represented using this model. Vector space model also does not consider the semantic relationships between attributes, which is very important in the email classification domain.

Paper [2] proposes a graph model to represent the concept in the sentence level to find document similarity. The concept follows a triplet representation. We use the same graph structure representation for our email text data.

3 Solution Approach

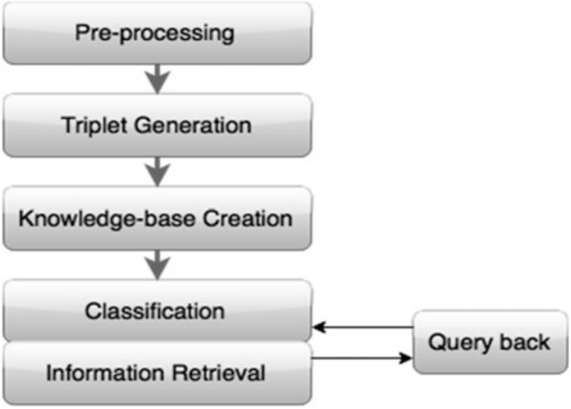
The proposed system will semantically classify the email queries received at service desks. The solution approach is shown in Fig. 1.

The solution methodology in this paper is divided into four modules. (1) Preprocessing (2) Triplet Generation (3) Knowledge base creation (4) Classification and Information Retrieval, which are described in Sects. 3.1–3.4.

3.1 Preprocessing

In this phase, Document Cleaning, Parts of Speech Tagging, and Phrase Structure Tree Generation are done as described in [2]. After the preprocessing phase, the verb argument structure is generated.

Fig. 1 Solution approach



3.2 Triplet Generation

The whole training data is represented as <S, V, O> Triplets, as described in paper [2], where S is the Subject, V is the Verb, and O is the Object. The triplet generation algorithm [2] is given below.

```
Algorithm 1 Triplet Generation Algorithm [1]
Input: A Document
Output: Concepts in the form of Triplets
S is a new sentence
Declare Lv as an empty list of Verb
Declare Sub as an empty list of Subject
Declare Obj as an empty list of Object
for each S do
    extract all verbs
    add verb to Lv
    for each verb in Lv do
        Check parent node and extract NP node
        Add NP node to Sub
    end for
    if verb contains NP or S as subtree then
        Add NP or S to Obj
    else if verb contains VP as subtree
        Add Object to Obj
    else
        Take parent node of verb node and search NP
    end if
end for
```

Example queries and corresponding triplets are given in Table 1.

3.3 Knowledge Base Creation

This knowledge base is the major module of the system. It is this knowledge base that we use for classification and information retrieval. The generated triplets from step B can be used to create the knowledge base. For that, we use the concept of graph data model [5]. The generated triplets are taken and a data graph is created using RDF format. Each query triplet will be converted to one RDF statement (triple). This data graph is the knowledge base, which serves as the basis of the

Table 1 Triplet generation

Query	Triplet
Windows has been expired in N200	<Windows, expired, N2011>
Scanner complaint at N200	<Scanner, complaint, N200>
Windows required at admin office	<Windows, required, Admin Office>

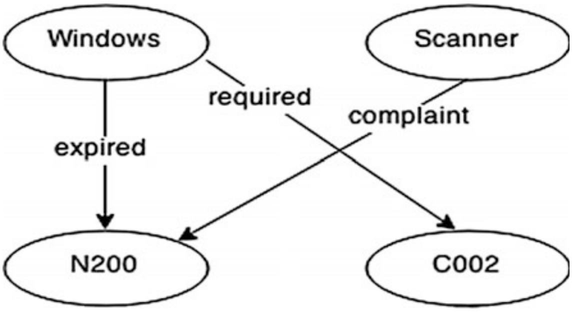


Fig. 2 Graph model

system. Also, we use OWL to define the semantic metadata of the RDF data, which will be explained in the following sub sections.

In graph model, text data can have arbitrary object relationships between each other. Graph model describes data using resources and relations. There will be resources, which will be related to other resources. A sample graph, which represents the query triplets in Table 1. As explained earlier, the complex relationships between the data can be noticed here.

Data in RDF are often called RDF Triples. An RDF Triple contains a Subject (Resource), a Predicate (Property), and an object (Resource). Each triple is called a statement also. Consider a single query from Fig. 2, for example, the triple *Windows* > *expired* > *N200.* It denotes a single query, "Windows has been expired in N200." Here, 'Windows' and 'N200' are resources, while 'expired' is a property. Though RDF is a way of describing data, it does not have a mechanism to define the semantics on its own. It does not say anything about 'Windows' or 'N200'. For this, we use OWL. OWL defines the semantics of the data described using RDF. We can create a hierarchy of those resources, and assign OWL classes for various resources and properties, which is often called taxonomy. For that, we use OWL classes, subclasses, individuals, and properties. What we create using OWL is the semantic metadata of our knowledge base, and it is called ontology. For example, given below is the sample ontology for our service desks query management (Fig. 3).

Here, 'Thing' is the root of our hierarchy. All OWL classes are the subclasses of 'Thing.' We can define any number of subclasses under 'Thing' class. The major subclasses present in our system are; 'Resource', 'Location', and 'Person.' We also define subclasses 'Software', 'Hardware', and 'Network' under 'Resource'. Under 'Location' we have subclasses like 'Classrooms', 'Departments', 'Labs', etc. 'Student', 'Faculty' and 'Other staff' are the subclasses defined under 'Person'. The items given in dotted lines, 'Windows', Wi-Fi', 'C001', 'L002', etc., are OWL 'Individuals'. They are instances of the OWL classes defined earlier.

Now the triple Windows > expired > N200 is more meaningful. Suppose that the above query is sent by the person with ID 'P003'. From the ontology, it is clear that 'P003' is a 'Faculty'. So, now the query tells the 'Software' item called 'Windows' has been expired in a 'Classroom' with number 'N200', and the query is sent by a

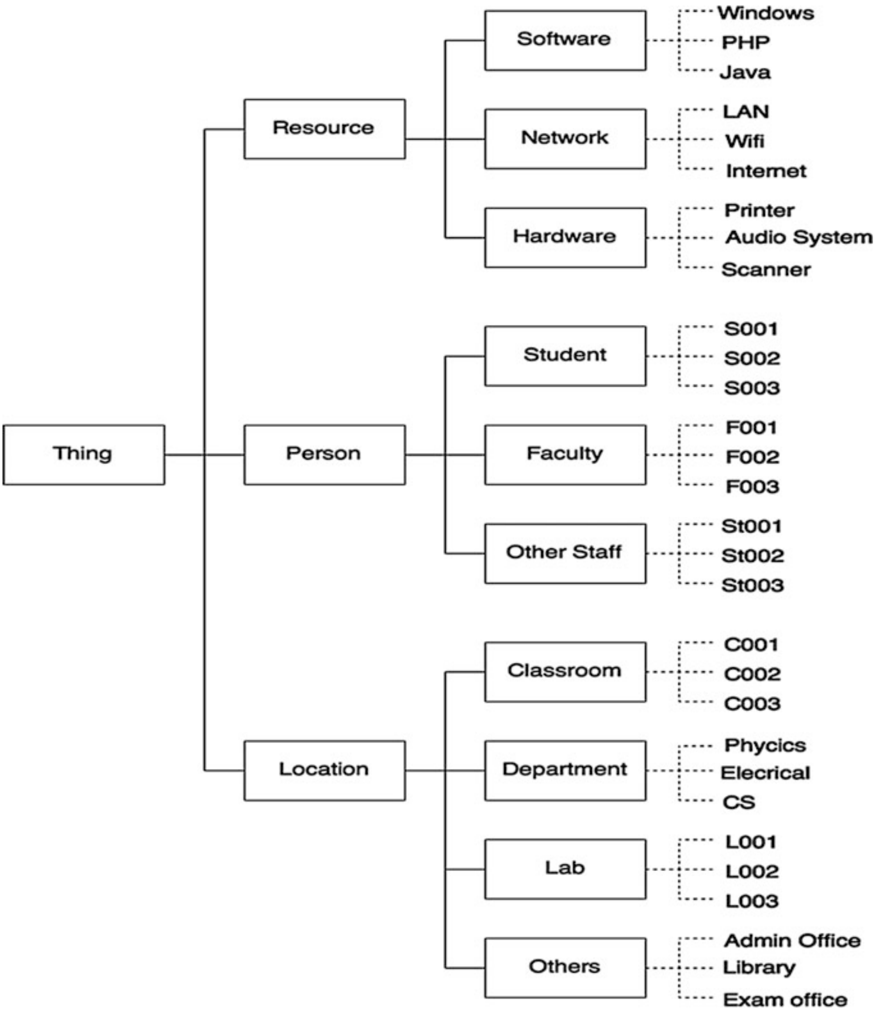


Fig. 3 Sample ontology of service desk query management

'Faculty', so it is a high priority query. Now that the knowledge base is defined, we can use it for Classification and Information Retrieval which are explained in the following sections.

3.4 *Classification and Information Retrieval*

Classification is one of the main aims of the system. In this work, we manage queries in the service desk of our university. In such service desks, usually the

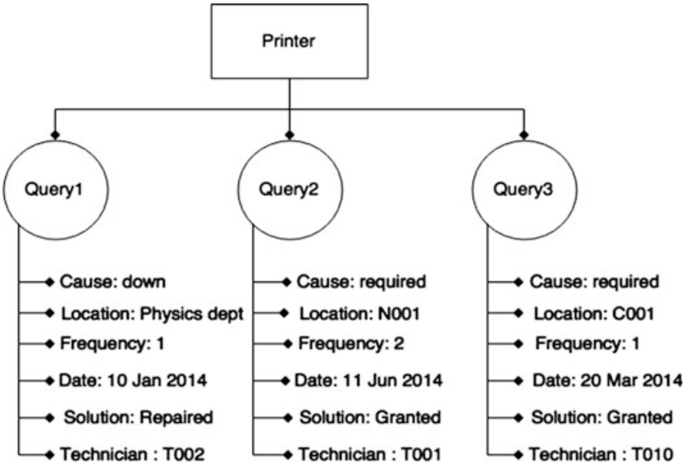


Fig. 4 Representation of training data

queries belong to ‘Software’, ‘Hardware’, or ‘Network’ category. So we have used these categories as subclasses in our ontology system. First we create a graph from the training data. All the queries are added to the graph. There can be multiple queries related to one subject. Each query is represented as a bag which contains the attributes related to the query. We define attributes like ‘Cause’ which denotes the nature of complaint, ‘Frequency’ which denotes how many times the particular query has occurred, ‘Date’ which denotes the date in which the query has occurred, ‘Solution’ which denotes the solution to the particular query and ‘Technician’ which denotes the details of the person who handled the query. For example, consider the following queries and their representation according to our system.

Queries: Printer Down at Physics Department, Printer Required in N001, Printer Required in C001

Figure 4 shows the representation of these queries in our system.

When a new query comes first convert it into graph form and the training data graph is searched for similar queries. If there are matching queries the attributes like ‘Solution’, ‘Frequency’, ‘Technician’, etc., can be retrieved to get useful information. The new query can be classified based on the subject, verb, or object, (or all of the three depending on the need) and it is assigned the most similar category. For example, if a query, “Printer down in L002” comes, first, it is converted into triplet form, i.e., <Printer, down, L002>. Now, according to the ontology, the query is classified as belonging to ‘Hardware’ category as well as it is classified as belonging to ‘Lab’ category indicating that the complaint occurred in one of the labs. Then the most similar query is “Printer down at Physics Department.” From the existing query it is clear that the technician ‘T001’ ‘repaired’ the printer. This result can be used for the current query also. The classification process is described in the algorithm given below.

Algorithm 2 Classification

```
T is the dataset
Q is the set of queries to be classified
C is the list which contains the output classes
for each query $q_i$ in Q do
    Generate triplets from $q_i$ ($q_i$ is a query sentence)
    S is the subject of the triplet
    V is the verb of the triplet
    O is the object of the triplet
        Create an RDF triple, R with S,V and O

    Search T for matching R
    if match found then
        get the class in $c_i$
        add $c_i$ to C
    end if
end for
```

The input to the algorithm is Q, the set of queries to be classified. After executing the algorithm, we will get the classified results in the list C.

Query Back and Information Retrieval

The user query may not always be fully informative. After examining the training data, we found that sometimes the query can be as vague as it cannot be classified to a category (e.g., of a vague query: '*System complaint*'). When these types of queries arrive for classification, a possible way is to query back the sender to get some more relevant information about the query. These are the steps that happen when a query that is very vague comes for classification. First the query is given to classification algorithm. The query will not get classified as it is too unclear. Then a reply message will be given to the sender, that the query is too vague, and additional information is needed to process it. This is done till the query is informative enough to get classified.

Effective information retrieval is another advantage of the system. As we have a large collection of training data in RDF documents, these documents can be queried to find out important information. RDF documents are queried using a query language SPARQL (SPARQL Protocol and RDF Query Language), which is a semantic language to query graph data.

When a new query comes, the knowledge base can be queried to extract some important information. For example, if a query comes, "Printer is required in C001," after triplet generation, we will get <Printer, required, C001>. By querying the knowledge base, we will get more information like what is 'Printer', what is 'C001', etc. It will find out that 'C001' is a conference room. Now, the location of 'C001' and other related attributes can be found from the ontology. Suppose that we are getting many more complaints for many other devices in the same place, and then we can infer that there is some problem with that place 'C001', and it may not be a problem with all the devices. It can be a power failure or something, which affected the place ‘C001’ as a whole. Again, if we query for the item ‘Printer’, all queries related to ‘Printer’ will be shown. So we can find the general issues with the item ‘Printer’.

4 Experiments and Results

All the experiments were conducted with the ICTS helpdesk email dataset. ICTS is the IT services provider of Amrita Vishwa Vidyapeetham University, Amritapuri. The dataset we used for experiments contain 72,000 emails queries, which were already labeled into three categories, namely, ‘Software’, ‘Hardware’, and ‘Network’. To represent the training data in graph model, we used the Apache Jena framework. For creating the ontology, a tool, ‘Protégé’ [6] is used, which offers a user-friendly GUI to create various hierarchies.

Results analysis is done based on the classification efficiency, which is calculated using *True Positives* (TP), *False Negatives* (FP), and *Precision* (P). True positives are the items correctly classified as belonging to the correct class while false positives are the ones which wrongly indicate that the item belongs to a particular class. Precision is the ratio of the number of true positives to the sum of the number of true positives and the number of false positives.

$$\text{Precision}, P = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{1}$$

The performance evaluation result of concept-based classification versus keyword-based classification is given below. Figure 5 illustrates the result analysis graph (Tables 2 and 3).

Fig. 5 Result analysis graph

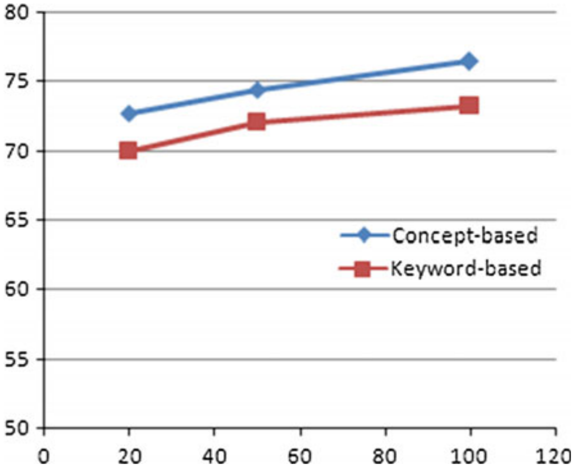


Table 2 Performance evaluation of concept-based classification

Performance measure	Number of queries		
	20	50	100
TP	16	40	78
FP	5	13	24
Precision (%)	72.72	75.47	76.47

Table 3 Performance evaluation of keyword-based classification

Performance measure	Number of queries		
	20	50	100
TP	14	38	74
FP	6	14	27
Precision (%)	70	73.07	74.26

In our system, we used concepts rather than keywords for creating the graph model. From the result, it is clear that concept-based classification offers higher precision than normal keyword-based approach.

Our advantage was, the types of devices which are registered with the service desk were already known. So, after creating the ontology of all known devices, the classification is found o be accurate. If a new device which is not already registered comes, it also can be added to the ontology so that queries related to that can also be classified accurately. It is found that when we add more Individuals to the ontology, the accuracy of classification increases.

5 Conclusion

Through this research work, an efficient way of service desk query classification and information retrieval has been implemented. Each query was represented, without losing semantics using triplets and modeled in a graph structure using RDF. Details about each type of device were stored in the ontology using OWL. From the knowledge base thus created, accurate query classification was done. Information retrieval from the knowledge base was done with the help of SPARQL query language.

Now, the system works only with structured sentences. The service desk queries can be unstructured also. So, our future work is to find an efficient way to represent unstructured query sentences and to manage them properly.

Acknowledgments We are thankful to Dr. M.R. Kaimal, Chairman, Department of Computer Science, Amrita Vishwa Vidyapeetham University, Amritapuri Campus for his valuable feedback and suggestions.

References

1. Salton, Gerard, Anita Wong, and Chung-Shu Yang. “A vector space model for automatic indexing.” *Communications of the ACM* 18.11 (1975): 613–620.
2. Veena, G., and N. K. Lekha. “A concept based clustering model for document similarity.” *Data Science & Engineering (ICDSE), 2014 International Conference on*. IEEE, 2014.
3. Leung, Nelson KY, and Sim Kim Lau. “Relieving the overloaded help desk: a knowledge management approach.” *Communications of the IIMA* 6.2 (2015): 11.
4. Rish, Irina. “An empirical study of the naive Bayes classifier.” IJCAI 2001 workshop on empirical methods in artificial intelligence. Vol. 3. No. 22. IBM New York, 200.
5. Chang, Jae-Yong, and Il-Min Kim. “Analysis and Evaluation of Current Graph-Based Text Mining Researches.” *Advanced Science and Technology Letters* 42 (2013): 100–103.
6. Gennari, John H., et al. “The evolution of Protégé: an environment for knowledge-based systems development.” *International Journal of Human-computer studies* 58.1 (2003): 89-123

Designing Chaotic Chirikov Map-Based Secure Hash Function

Shruti Khurana and Musheer Ahmad

Abstract In this paper, a new algorithm for designing one-way cryptographic hash function using chaotic Chirikov map is proposed. High sensitivity of Chirikov system makes it an ideal candidate for secure hash function design. In the proposed algorithm, the input message is split into a number of small blocks. The blocks are processed using chaotic map. The two intermediate hashes are generated using evolved control and input parameters. The two intermediate hash values are then employed to yield the final variable-length hash. The simulation and statistical analyses illustrate that the anticipated hash algorithm exhibits encouraging lineaments like high sensitivity to input messages and key, satisfactory confusion and diffusion attributes which verify the suitableness of proposed algorithm for the design of secure cryptographic variable-length hash functions.

Keywords Hash function · Chaotic Chirikov map · Security · Confusion and diffusion

1 Introduction

Cryptographic hash functions are a kind of unique one-way functions which take as input messages of variable length and give an output of defined length. Secure hash functions may be categorized as types: *unkeyed* hash functions, which take only an input parameter, i.e. a message in its specification; and *keyed* hash functions, which include usually input message and secret key, i.e. the generated hash is under the control of message and the key as well [1, 2]. Since the generated hash is dependent

S. Khurana (✉) · M. Ahmad
Department of Computer Engineering, Faculty of Engineering and Technology,
Jamia Millia Islamia, New Delhi 110025, India
e-mail: s1994khurana@gmail.com

M. Ahmad
e-mail: musheer.cse@gmail.com

H.S. Saini et al. (eds.), *Innovations in Computer Science and Engineering*,
Advances in Intelligent Systems and Computing 413,
DOI 10.1007/978-981-10-0419-3_32

on all parts of message at bits, characters or block levels, any minute alteration in one bit or character of message must result in drastic capricious effects in generated hash. Applications of hash functions include detection of errors in file transfer, generation of MAC (message authentication code), password storage, etc. [3]. Examples of some conventional number-theoretic hashes include MD5, SHA-1, SHA-2 and SHA-3 [4]. Researchers have investigated that hash function algorithms like MD5 and SHA-1 are no longer secured and broken. The collision attack on the existing algorithms such as MD5 and SHA-1 is either very easy to carry out or is very close to practical [4–6]. Therefore, it is imperative to come up with effective means for establishing strong hashes. The characteristics of secure hash codes include the following [7, 8]:

1. The input message may take any bit length without any constraints, but its output must be of fixed bit length.
2. Hash functions must be irreversible, which implies that for all known outputs, it is inconceivable to predict any part of message which hashes to that corresponding output.
3. It is impracticable to have other input which has exact output hash as the first input.
4. No two distinct inputs x, x' can produce a hash $h(x)$, $h(x')$ such that $h(x)$, $h(x')$ are the same, i.e. $h(x') = h(x)$ is not possible.

Of late, a number of chaotic maps based cryptographic hash schemes have been suggested in the literature. In the chaos theory, chaotic systems deal with non-linear dynamic systems, these systems are extremely sensible to initial conditions and exhibit deterministic random-like operation. Any small change in initial condition could result in vast difference in the final outcome. Apart from sensitivity to minor alterations in initial conditions and parameters, other cryptographically suited features of chaotic sequences include: random-like behaviour, topological mixing, which collectively provides and fulfils the requirements of confusion and diffusion needed for any cryptographic process. The researchers have studied the analogy between the chaos and cryptographic properties [9]. In the past two decades, researchers have applied extensively the chaotic systems to design strong cryptographic primitives like data encryption, data hiding, hash functions, algorithms, etc. The Chirikov map is also known as the standard map. It is an area-preserving system bounded within a square of side 2π. The Chirikov map is constructed by a Poincaré's surface of section of the kicked rotator. The chaotic behaviour of this map was manifested by Boris Chirikov [10]. The Chirikov map governed by the system is given as

$$\begin{aligned} x(k+1) &= (x(k) \ + a\sin(y(k)))\ \mathrm{mod}\ (2\pi) \\ y(k+1) &= (y(k) \ + x(k+1))\ \mathrm{mod}\ (2\pi) \end{aligned}$$

Here, a novel hash code generation function is proposed using Chirikov map and the effectiveness of the algorithm is investigated statistically. The proposed algorithm generates a secure hash function which not only satisfies the security requirements, but also has ample hash functions confusion and diffusion.

The placement of the remainder of the paper is organized as follows: Sect. 2 devotes to the explication of proposed chaos-based hash function algorithm. The cryptographic strength of proposed hash algorithm is quantified through statistical parameters and examined in Sect. 3. The conclusion and summary of proposed work is made in Sect. 4.

2 Proposed Hash Function

The algorithmic steps of proposed hash generation method are furnished as below.

H.1. Let input message be $S_{n\times 1}$ where n is characters in St.
H.2. Provide initial conditions of chaotic map as x_0, y_0, a.
H.3. Iterate Chirikov map 10 times and discard the values.
H.4. Decompose message S into blocks, each of size 8 characters.
H.5. l = floor(n/8)
H.6. $i = 1$
H.7. Repeat Steps 8–10 for all blocks of message while $i \leq l*8$.
H.8. Further iterate standard map once.
H.9. Find control parameters: $N_1 = N_1 + f_1\ (S_{i+1 \text{ to } i+4})$ and $N_2 = N_2 + f_2\ (S_{i+5 \text{ to } i+8})$.
H.10. Increment i by 8.
H.11. Calculate $r = n - (l * 8)$.
H.12. Find the remaining message characters (if any) = $S_{l*8 \text{ to } n}$
H.13. If $r = 1$, calculate: $N_1 = N_1 + f_3\ (S_{l*8+1})$ and $N_2 = N_2 + f_4\ (S_{l*8+1})$.
H.14. If $r = 2$, calculate: $N_1 = N_1 + f_5\ (S_{-l*8+1})$ and $N_2 = N_2 + f_6\ (S_{-l*8+2})$.
H.15. If $r = 3$, calculate: $N_1 = N_1 + f_7\ (S_{l*8+1 \text{ to } l*8+2})$ and $N_2 = N_2 + f_8\ (S_{l*8+2 \text{ to } l*8+3})$.
H.16. If $r = 4$, calculate: $N_1 = N_1 + f_9\ (S_{l*8+1 \text{ to } l*8+2})$ and $N_2 = N_2 + f_{10}\ (S_{l*8+3 \text{ to } l*8+4})$.
H.17. If $r = 5$, calculate: $N_1 = N_1 + f_{11}\ (S_{l*8+1 \text{ to } l*8+3})$ and $N_2 = N_2 + f_{12}\ (S_{l*8+3 \text{ to } l*8+5})$.
H.18. If $r = 6$, calculate: $N_1 = N_1 + f_{13}\ (S_{l*8+1 \text{ to } l*8+3})$ and $N_2 = N_2 + f_{14}\ (S_{l*8+3 \text{ to } l*8+6})$.
H.19. If $r = 7$, calculate: $N_1 = N_1 + f_{15}\ (S_{l*8+1 \text{ to } l*8+4})$ and $N_2 = N_2 + f_{16}\ (S_{l*8+4 \text{ to } l*8+7})$.
H.20. Repeat Steps 21–23 for k = 1 to cnt times, where cnt = hash_length/8.
H.21. Further iterate Chirikov map once.
H.22. Generate a as:

$$a = \left(\frac{N_1}{N_2}\right) \times \left(\frac{\text{floor}(y \times (10^{10}) \bmod (2^{20}))}{\text{floor}(x \times (10^{10}) \bmod (2^{20}))}\right) \times a$$

H.23. Generate hash character as: $H_{\text{cnt}} = f_{17}\ (x, y)$.
H.24. Store the hash generated in Steps 21–23 as HASH1.
H.25. Proceed further with current input and control parameters. Repeat Steps 3–23 to generate another hash HASH2.
H.26. Repeat Step 27 for k = 1 to cnt times.

H.27. FHASH$_k$ = f_{18} (HASH1$_k$, HASH2$_k$).
H.28. Output the FHASH as final hash.

Various functions and symbols used in the algorithm are listed below. In functions f_1 to $f_{18,}$ S_i refers to the value of input message character '*i*' taken as ASCII integer.

f_1 $(S_{i+1 \text{ to } i+4}) = S_{i+1} + S_{i+2} + S_{i+3} + S_{i+4}$
f_2 $(S_{i+5 \text{ to } i+8}) = S_{i+5} + S_{i+6} + S_{i+7} + S_{i+8}$
f_3 $(S_{l*8+1}) = S_{l*8+1}$
f_4 $(S_{l*8+1}) = S_{l*8+1} + S_{l*8+1}$
f_5 $(S_{l*8+1}) = S_{l*8+1}$
f_6 $(S_{l*8+2}) = S_{l*8+2}$
f_7 $(S_{l*8+1 \text{ to } l*8+2}) = S_{l*8+1} + S_{l*8+2}$
f_8 $(S_{l*8+2 \text{ to } l*8+3}) = S_{l*8+2} + S_{l*8+3}$
f_9 $(S_{l*8+1 \text{ to } l*8+2}) = S_{l*8+1} + S_{l*8+2}$
f_{10} $(S_{l*8+3 \text{ to } l*8+4}) = S_{l*8+3} + S_{l*8+4}$
f_{11} $(S_{l*8+1 \text{ to } l*8+3}) = S_{l*8+1} + S_{l*8+2} + S_{l*8+3}$
f_{12} $(S_{l*8+3 \text{ to } l*8+5}) = S_{l*8+3} + S_{l*8+4} + S_{l*8+5}$
f_{13} $(S\ _{l*8+1 \text{ to } l*8+3}) = S_{l*8+1} + S_{l*8+2} + S_{l*8+3}$
f_{14} $(S\ _{l*8+4 \text{ to } l*8+6}) = S_{l*8+4} + S_{l*8+5} + S_{l*8+6}$
f_{15} $(S_{l*8+1 \text{ to } l*8+4}) = S_{l*8+1} + S_{l*8+2} + S_{l*8+3} + S_{l*8+4}$
f_{16} $(S_{l*8+4 \text{ to } l*8+7}) = S_{l*8+4} + S_{l*8+5} + S_{l*8+6} + S_{l*8+7}$
f_{17} (x, y) = (floor $(x * 10^{10})$ mod (255)) XOR (floor $(y * 10^{10})$ mod (255))
f_{18} (HASH1$_{k,}$ HASH2$_k$) = (HASH1$_k$ AND HASH2$_x$) XOR (HASH1$_k$ OR HASH2$_k$)

3 Performance Evaluation

3.1 Sensitivity of Message

The standard procedure used in Refs. [11, 12] is adopted to evaluate and examine the performance of a cryptographic hash function generation method. Initially, an input message having 1024 null characters is picked. To demonstrate the sensitivity of message to the generated hash code comprehensively, the following illustration is acquired.

Condition 1: The original message contains 1024 null characters.
Condition 2: Change last character to 1.
Condition 3: Add one bit to the last character of the message.
Condition 4: Add a space to the end of the message.
Condition 5: Change a = 30.4 to 30.4000000001.
Condition 6: Change x_0 = 0.25 to 0.25000000001.
Condition 7: Change y_0 = 0.50 to 0.50000000001.

The corresponding 128-bit hashes in hexadecimals are handed as:

Condition 1: BDA258AF1D1FBA6AE008FF30FCBAB9E7
Condition 2: B717D91E47A542960EE3AE9C47E6DDF1
Condition 3: 73D53D0C0752D681BAA12564513458B0
Condition 4: F7E0BE4DAE5C99675AB4523547A829F4
Condition 5: C1A744522688BE6AFC1ACBB98E3F7D52
Condition 6: B6334F0ABFBEA92D72AE8374756780CC
Condition 7: EE2BBC26E140017A0FAAECC9D8F4BDD6

The above hash results contend that the requisite one-way attribute is fully satisfied and the minute alteration in plaintext message or key value causes immense adjustments in final hashes.

3.2 Statistical Analysis

Claude Shannon ascertains it is possible to assess the security performance of various kinds of ciphers and security primitives through statistical analysis [13]. He recommended the confusion and diffusion properties for the purpose of testing hash algorithms to mitigate statistical attacks. In an ideal diffusion effect, tiny changes in the initial conditions should have a probability of 50 % of changing each bit of hash. An input message of any size is taken and then its corresponding hash is produced; then, a single bit of message is altered indiscriminately and a fresh corresponding hash is yielded. The two hash codes are equated to one another, and the altered bits are accounted and called B_i. The above procedure is performed N times. The statistical measures used to quantify the Shannon properties are outlined mathematically below (Table 1):

Minimum altered bit number $B_{\min} = \min\left(\{B_i\}_1^N\right)$

Maximum altered bit number $B_{\max} = \max\left(\{B_i\}_1^N\right)$

Mean altered bit number $\overline{B} = \sum_1^N \frac{B_i}{N}$

Mean altered probability $P = \frac{B}{128} \times 100\ \%$

Table 1 Statistical number of altered bits B_i for $N = 512$, 1024, 2048 and 10,000 for 128-bit hash algorithm investigated in Ref. [12]

Parameters	In Ref. [12]			
	512	1024	2048	10,000
B	63.8808	63.8339	63.8945	63.9192
P (%)	49.9069	49.8703	49.9176	49.9369
ΔB	6.0032	5.8662	5.7711	5.6467
ΔP (%)	4.6900	4.5830	4.5087	4.4115
$B_{\min}$	45	45	43	41
$B_{\max}$	80	82	82	84

Table 2 Statistical number of altered bits B_i for $N = 256$, 512, 1024, 2048 and 10,000 for proposed 128-bit hash algorithm

Parameters	In proposed				
	256	512	1024	2048	10,000
B	63.9375	63.9433	64.2675	64.0356	64.0668
P (%)	49.9511	49.9557	50.2090	50.0278	50.0521
ΔB	5.8036	5.6671	5.6104	5.7599	5.6238
ΔP (%)	4.5341	4.4274	4.3831	4.4999	4.3936
B_{min}	46	48	46	46	40
B_{max}	78	80	84	83	86

Table 3 Statistical number of altered bits B_i for $N = 256$, 512, 1024, 2048 and 10,000 for proposed 160-bit hash algorithm

Parameters	In proposed				
	256	512	1024	2048	10,000
B	80.0	79.9042	80.0019	80.0712	80.0788
P (%)	50.0	49.9401	50.0012	50.0445	50.0492
ΔB	6.3239	6.4535	6.1705	6.2798	6.2963
ΔP (%)	3.9524	4.0334	3.8566	3.9249	3.9352
B_{min}	61	59	58	58	59
B_{max}	98	96	102	102	102

$$\text{Standard variance of the altered bit number } \Delta B = \sqrt{\frac{1}{N-1}\sum_{1}^{N}\left(B_i - \overline{B}\right)^2}$$

$$\text{Standard variance of probability } \Delta P = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(\frac{B_i}{128} - P\right)^2} \times 100\ \%$$

The statistical results of tests for $N = 256$, 512, 1024, 2048 and 10,000 are listed for 128-bit hashes in Table 2. It is evident from the resultant statistical outcomes that both average altered bit number B and average altered probability P are tremendously near to idealistic scores such as 64-bits and 50 %. Also, ΔB as well as ΔP are minuscule, evidencing that the capableness for diffusion and confusion is really strong and stable. The results obtained are consistent comparable to other chaos-based hashes recently investigated in [12] (Table 3).

3.3 Resistance to Birthday Attack

According to the Birthday attack problem, if there are 'N' different possibilities of something, then you need square root of 'N' randomly chosen items in order to have a 50 % chance of collision. The classical Birthday attack is applied to crack hash functions [3, 6]. Thus for hash function of 64-bit length, the attack complexity is not 2^{64}, but it is only 2^{32}. So, there is a chance of 50 % collision in only 2^{32}

attempts. Taking into account the computing ability of modern systems, the hash length should have a minimum length of 128-bits, thereby making the attack difficulty quite complex. The proposed hash size is 128-bits, and can be easily expanded to any greater length like 160, 256, 512, etc., with minimal alteration to the algorithm.

3.4 Flexibility

The suggested hash code generation algorithm is suggested with an aim to alleviate the issues like hash functions size, input message padding, immunity against brute force attack or birthday attack, etc. This algorithm is prepared such that no extra bits are needed along with the original input message. The algorithm is designed as a key dependent hash function as the initial values of Chirikov map served as keys. The established hash algorithms like MD5, MACDES and HMAC-MD5 have pre-specified sized output hash code. But the proposed algorithm is beneficial in terms that it can be used to generate hash of any length such as 160, 256, 512 and 1024 bits, by making a little adjustment in scheme. Furthermore, the anticipated algorithm is implemented in double-precision floating-point arithmetic environment.

4 Conclusion

In this paper, we proposed to present a new hash function method based on chaotic Chirikov map. The chaotic Chirikov map provides necessary sensitivity to the message such that even a minute change in secret key brings drastic changes in the hash code generated. The one-way attribute of cryptographic hashes is also ensured. The proposed hash function supports many features such as security requirements, having average altered probability close to the idealistic score of 50 %, stability, consistency and strong statistical diffusion and confusion capabilities. Furthermore, the anticipated hash algorithm provides the flexibility to expand the hash length to any arbitrary length, usually a multiple of 2. Besides, the proposed algorithm has computational simplicity.

References

1. Lian, S., Sun, J., Wang, Z.: Secure hash function based on neural network. Neurocomputing, 69(16), 2346–2350 (2006).
2. Wong, K. W.: A combined chaotic cryptographic and hashing scheme. Physics Letters A, 307 (5–6), 292–298 (2003).
3. Menezes, A. J., Oorschot, P. C. V., Vanstone, S. A.: Handbook of applied cryptography, CRC Press (1997).

4. Secure Hash Standard. Federal Information Processing Standards Publications (FIPS PUBS), 180(2), (2002).
5. Singla, P., Sachdeva, P., Ahmad, M.: Exploring Chaotic Neural Network for Cryptographic Hash Function. Emerging Trends in Computing and Communication. Springer India, LNEE 143–148 (2014).
6. Wang, X., Yin, Y. L., Yu, H.: Finding collisions in the full SHA-1. Lecture Notes in Computer Science, 3621, 17–36 (2005).
7. Xiao, D., Liao, X., Wang, Y.: Parallel keyed hash function construction based on chaotic neural network. Neurocomputing, 72(10), 2288–2296 (2009).
8. Yang, H., Wong, K. W., Liao, X., Wang, Y., Yang, D.: One-way hash function construction based on chaotic map network. Chaos, Solitons & Fractals, 41, 2566–2574 (2009).
9. Li, S., Chen, G., Mou, X.: On the dynamical degradation of digital piecewise linear chaotic maps. International Journal of Bifurcation and Chaos, 15(10), 3119–3151 (2005).
10. Chirikov, B.V.: Research concerning the theory of nonlinear resonance and stochasticity. Preprint N 267, Institute of Nuclear Physics, Novosibirsk (1969) (http://www.quantware.ups-tlse.fr/chirikov/refs/chi1969e.pdf).
11. Akhshani, A., Behnia, S., Akhavan, A., Jafarizadeh, M.A., Hassan, H., Hassan, Z.: Hash function based on hierarchy of 2D piecewise nonlinear chaotic maps. Chaos, Solitons & Fractals, 42(4), 2405–2412 (2009).
12. Akhavan, A., Samsudin, A., Akhshani, A.: A novel parallel hash function based on 3D chaotic map. EURASIP Journal on Advances in Signal Processing, 2013(126), (2013).
13. Shannon, C. E.: Communication theory of secrecy systems. Bell Systems Technical Journal, 28, 656–715 (1949).

Data Fusion Approach for Enhanced Anomaly Detection

R. Ravinder Reddy, Y. Ramadevi and K.V.N. Sunitha

Abstract Anomaly detection is very sensitive for the data because, the feature vector selection is a very influential aspect in the anomaly detection rate and performance of the system. In this paper, we are trying to revise the dataset based on the rough genetic approach. This method improves the quality of the dataset based on the selection of valid input records to enhance the anomaly detection rate. We used rough sets for pre-processing the data and dimensionality reductions. Genetic algorithm is used to select proper feature vectors based on the fitness. The fusion of the soft computing techniques improves the data quality and reduces dimensionality. Empirical results prove that it improves detection rate as well as detection speed.

Keywords Anomaly detection · Rough set theory · Genetic algorithm · Feature selection · Classification

1 Introduction

Usage of internet has increased enormously; with this, everyone is using and benefitting the services of the Internet. Exchanging the information via Internet is common these days than the other modes of communication. Along with it, the

R. Ravinder Reddy (✉) · Y. Ramadevi
Department of Computer Science and Engineering,
Chaitanya Bharathi Institute of Technology, Hyderabad 500075, India
e-mail: ravindra_rkk@cbit.ac.in

Y. Ramadevi
e-mail: yrd@cbit.ac.in

K.V.N. Sunitha
B.V. Raju Institute of Technology for Women, Bachupally,
Hyderabad 500090, Telangana, India
e-mail: k.v.n.sunitha@gmail.com

H.S. Saini et al. (eds.), *Innovations in Computer Science and Engineering*,
Advances in Intelligent Systems and Computing 413,
DOI 10.1007/978-981-10-0419-3_33

threat to the information also increased gradually. Intruders are trying to get this information for doing an unauthorized transaction in the web. Many techniques are available to find the intruder's actions in the system. Most of the people are applying machine learning, data mining [1, 2] and soft computing techniques to detect the intruder's attitude.

Intrusion detection is an attempt to protect the system resources and detect threats which can compromise the security triangle C.I A. It acts like a second wall of security to the system. Firewall has several limitations when compared with intrusion detection system; it prevents some information coming from un-trusted network and limits the access. It provides some sort of physical security, unable to protect the threats from insiders.

From the inception model of the intrusion detection by Denning [3], populous researchers have applied different technologies to identify the intrusion behaviour including machine learning, data mining and soft computing techniques. All these methods endeavour to improve the detection rate and reduction in false alarm rate significantly compared to the Denning's inception model. Identifying the intrusion detection is basically treated as a classification [2] problem which will classify the anomaly activity from normal behaviours. Choosing appropriate classifier is an important task in the determining of anomaly behaviour. The classifier performance is directly dependent on the quality of training data, for improving the data quality here, we proposed a hybrid approach and is called rough genetic method. It extracts the quality feature vectors from the given population.

In these days, soft computing is applied for almost all the fields like that and we used this model for enhancing the anomaly detection. In recent trends, combination of soft computing techniques are applied to intrusion detection for increasing the detection rate. In this paper, we combined the rough set theory for dimensionality reduction and feature selection. Once optimal feature set is obtained, Genetic algorithm is applied to extract the feature vector from the available population.

Genetic Algorithm [4] searches the best solution from all the possible solutions. GA has been applied for best searching technique for extracting suitable features from the initial population. In this method, for generating new population we used the genetic algorithms. In this paper, we used these genetic operators include reproduction, crossover and mutation, dropping condition. Fitness functions ensure that the evolution is toward optimization by calculating the fitness value for each individual in the population. Here, the dataset has trained to develop the genes for the evolution of the new fitness population.

The remaining of the paper is organized as follows, in Sect. 2 we brief the concepts used in this paper. In Sect. 3 the methodology, experimentation and results are discussed along with Sect. 4. Finally the conclusion and future work in Sect. 5.

2 Related Work

In this section, brief outline of the concepts are presented.

2.1 Intrusion Detection

Intrusion is an attempt to access the system resources in an unauthorized way to modify or destroy the resources from outsiders or may be the insiders. So, intrusion detection system is the second wall of the protection of the system. Basically based on the behaviour of intruders it divides into two aspects.

A. **Misuse detection/signature-based**

In this approach, user has to define the predefined patterns or signatures to detect the malicious behaviour. These types of intrusion detection systems can be called as statistical systems like snort, Bro, etc. These systems will detect the known attacks accurately but the problem with this is that it would not detect the unknown attacks.

B. **Anomaly Detection**

Intruder's behaviour is dynamic in nature; to break the security firewalls they will come up with new patterns of attacks in disguised manner. In these situations, we need a dynamic model to detect the attacks. Anomaly detection is used to detect these types of attacks in dynamic nature. It will detect novel attacks but the false alarm rate is high.

Based on the type of data intrusion detection can be divides into three categories.

A. **Host-Based IDS**

These data come from the records of different host system activities, system logs and application program statistics.

B. **Network-Based IDS (NIDS)**

NIDS collects data from the network devices like sensors, routers and ports for analyzing the network connection records. They examine the network stream of traffic and check whether it falls within acceptable boundaries.

C. **Distributed IDS**

This type of system will collect the data from different agents and analyze for anomaly behaviour.

2.2 Rough Set Theory (RST)

Rough set theory introduced by Pawlak [5–7], is a mathematical tool used for representing an imprecise and vague data. Recent research is focused on RST-based machine learning and data mining. It is based on the approximations; in this method, it will calculate the lower and upper boundaries. If the difference of these boundaries is null then it is crisp set otherwise Rough set. It works on approximations, the concepts mainly used for optimal feature selection from the given data.

Compared to the other feature selection techniques RST has a solid mathematical framework is established to model relationships between attributes with a minimal rule set. Finding optimal features in large information systems is still an NP-hard problem [8, 9].

Compared with conventional techniques, Rough Set Theory has the following advantages:

(1) The model will learn from the training datasets of small size. RST provides a systematic method capable of searching and identifying the relationships within the data attributes of a relational database.
(2) Simplicity: Generally, the simpler the model the higher the detection efficiency. RST provides a systematic method to obtain a set of rules with a minimal size. This makes the rules extracted by RST suitable for real-time detection tasks.

2.3 Dataset

To evaluate any system, we need a benchmark input and compare the results. Fortunately, for evaluation of the intrusion detection system, we have The KDDCUP'99 dataset [10, 11]. Since 1999, KDDCUP'99 has been the most wildly used dataset for the evaluation of anomaly detection methods. In the feature vector, all attributes may not be critical to the evaluation of intrusion detection. It is a public repository to promote the research works in the field of intrusion detection. It contains 41 conditional attributes and one class label. It is a standard dataset being used for intrusion detection. In this, the attacks are distributed in a probabilistic manner.

2.4 Genetic Algorithm (GA)

Soft computing evolution has changed a lot of difference in the computing era. Among all, the Genetic Algorithms (GA) is playing crucial roles in the security environment. GA encodes potential solutions for a specific problem; it generates the new chromosomes from the existing population by applying the operators with the defined fitness for survival. These newly generated chromosomes with survival fitness will perform better than the existing [12–14]. Using the operators and fitness generating next population is called a generation.

GA's are adaptive heuristic search algorithms used to solve optimization problems. Basically, it works on the principle of "survival of the fittest" laid by Darwin. In nature, competition among individuals for scanty resources results in the fittest individuals dominating over the weaker ones. Generating the fittest individual from the existing population requires more number of iteration; it search for the fittest value. GA's will give robust solutions; this is the main advantage of it. We adopt this revolutionary approach for detecting novel intrusion as well as to improve detection rate.

A GA first defines the following to perform further operations:

1. A genetic representation of the solution domain,
2. A fitness function to evaluate the solution domain.

Every generation, individuals are replaced with the new ones by the following genetic operators in order to obtain the maximum accuracy of the classifier.

1. Selection
2. Crossover
3. Mutation

3 Methodology

Most of the existing Intrusion detection systems suffer with the detection rate and time taken to detect the intrusion. In this paper, the proposed method handles these issues attentively. We divide the task into two modules as follows.

1. Input data selection
2. Classification

As the process shown in Fig. 1, the KDDCUP99 dataset is given to the data fusion module, in this pre-process the data. Discretization is the prerequisite to the rough set theory [11]; we performed discretization first and calculated the reduct using the quick reduct algorithm.

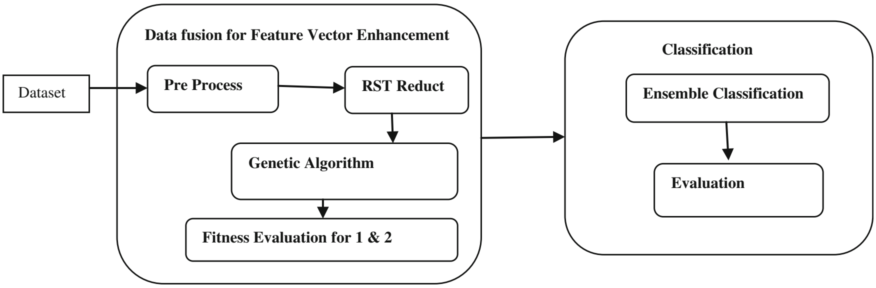


Fig. 1 Data fusion approach for IDS model

A Quick reduct algorithm as follows:

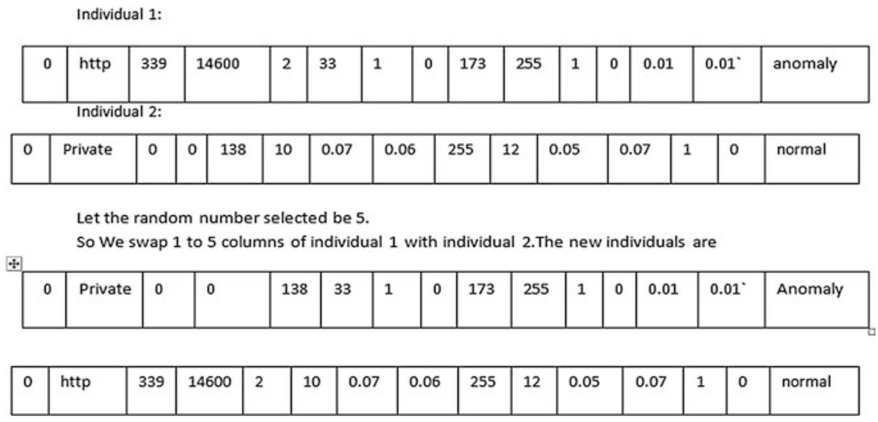
```
**Algorithm**: QuickReduct(C D R)
Input: The set C of all conditional attributes
The set D of decision attributes.
Output: The reduct R of C($R \subseteq C$)
1. $R \leftarrow \Phi$
2. do
3. $T \leftarrow R$
4. $\forall x \in (C-R)$
5. if $\gamma R \cup \{x\}$ (D) $> \gamma T(D)$
6. $T \leftarrow R \cup \{x\}$
7. $R \leftarrow T$
8. until $\gamma R(D) = \gamma C(D)$
9. return R
```

The quick reduct algorithm reduces the dimensionality of the feature vector size from 42 to 15 attributes. It saves the computational space and time. The optimized feature vector is used to calculate the new population using the genetic algorithm. As explained previously, our genetic algorithm contains three important operators. The important input for this is choosing the right fitness value.

3.1 Crossover

We randomly select a part of the population and swap it in the next population, in Fig. 2 is shown the sample crossover operation for the given feature vectors.

In the similar way, we repeat it for every individual from the initial population.



Individual 1:

0	http	339	14600	2	33	1	0	173	255	1	0	0.01	0.01`	anomaly

Individual 2:

0	Private	0	0	138	10	0.07	0.06	255	12	0.05	0.07	1	0	normal

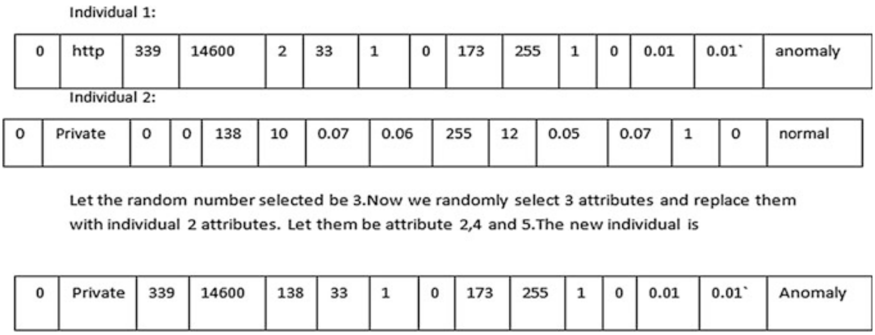
Let the random number selected be 5.

So We swap 1 to 5 columns of individual 1 with individual 2.The new individuals are

0	Private	0	0	138	33	1	0	173	255	1	0	0.01	0.01`	Anomaly

0	http	339	14600	2	10	0.07	0.06	255	12	0.05	0.07	1	0	normal

Fig. 2 Crossover operation



Individual 1:

0	http	339	14600	2	33	1	0	173	255	1	0	0.01	0.01`	anomaly

Individual 2:

0	Private	0	0	138	10	0.07	0.06	255	12	0.05	0.07	1	0	normal

Let the random number selected be 3.Now we randomly select 3 attributes and replace them with individual 2 attributes. Let them be attribute 2,4 and 5.The new individual is

0	Private	339	14600	138	33	1	0	173	255	1	0	0.01	0.01`	Anomaly

Fig. 3 Mutation of feature vector

3.2 Mutation

Mutation operation is shown in Fig. 3 for the feature vector. After the mutation, we generate a new feature vector.

Hence, we do it for all iterations. Now, we calculate the fitness for these new populations which generated newly. Based on the fitness we select the new population. Here, we are evaluating for the two different fitness values, fitness 1 and 2. These steps will repeated till the generation of new population of n-feature vectors with the required fitness.

3.3 Fitness

The fitness value [4] evaluates the performance of each individual in the population. We use a fitness function defined based on the support–confidence framework. Support is a ratio of the number of records covered by the rules to the total number of records [15]. Confidence factor (cf) represents the accuracy of rules, within the confidence interval how many will be true under this rule. It is the ratio of the number of records matching both the consequent and the conditions to the number of records matching only the conditions.

Once the new population is generated with data fusion, ensemble classifier is used to classify the anomaly behaviour. The ensemble technique is a combination of classifier for enhancing the classifier performance. It collects the opinion of the classifiers and reduces the bias and variance. It handles imbalanced class problems in the data. Here, we used bagging technique for ensemble classifier. The Bagging [16] algorithm creates an ensemble of classification techniques; each learning scheme gives an equally weighted prediction.

```
Algorithm: Bagging
Input:  D, a set of d training tuples.
        k, the number of models in the ensemble, a classification learning scheme.
Output: The ensemble a composite model, M*;
Method:
    for i=1 to k do    // Create models
        Create bootstrap sample, Di , by sampling D with replacement;
        Use Di and the learning scheme to derive a model, Mi;
    end for
```

To Use the ensemble to classify a tuple X, let each of the k models classify X and return the majority vote.

4 Experiments and Results

The working of each module is described as follows:

We have conducted the experimentation for the KDDCUP'99 dataset. As depicted in Fig. 4 the processes flow is as follows.

Here, the experimentation is conducted for the KDDCUP'99 dataset. First, pre-process the dataset. Then, we applied the rough set approach for the feature selection and dimensionality reduction. GA is applied for calculating the new population based on the fitness function. The feature vectors that are satisfying the fitness will be selected for the new population. We generated few records based on

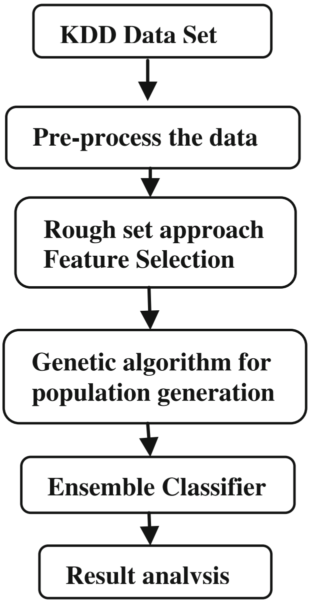

KDD Data Set
Pre-process the data
Rough set approach Feature Selection
Genetic algorithm for population generation
Ensemble Classifier
Result analvsis

Fig. 4 Processes flow of the system

the genetic operations for satisfying the fitness value. Once the dataset is improved in quality and reduced the dimensionality, with this the system performance has increased enormously.

Algorithm The Rough Genetic NIDS
Input KDD Cup dataset
Output The anomaly detection rate

1. Pre-process the dataset to the required format
2. Reduce the Dataset using Quick Reduct Algorithm.
3. Using the genetic algorithm, calculate the feature set based on the given fitness value
4. Once the tuples are generated by GA, apply the rough set approach for feature selection
5. Apply the ensemble classifier to the updated dataset.

In this process, we achieved a better detection rate as well as increased the detection speed. This process can be used in real environments to detect the anomaly behaviour of the system. The accuracy of an intrusion detection system is measured regarding the detection rate and false alarm rate. These may not be sufficient for evaluating a system performance. The further extensions of these measures combine and give the detailed quality measures to the system.

$$\begin{aligned} \text{Accuracy} &= \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \\ \text{Precision (Effectiveness)} &= \frac{\text{TP}}{\text{TP} + \text{TN}} \\ \text{Recall (Ability)} &= \frac{\text{TP}}{\text{TP} + \text{FP}} \\ \text{F-measure} &= \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \end{aligned}$$

The performance measures of the given system are shown in Table 1, it shows an enhanced results.

In Figs. 5 and 6 are the plotted graphs for different measures. It shows the significant performance improvements.

Table 1 Performance for the different fitness values

Fitness	TP rate	FP rate	Precision	Recall	F-measure
1	0.98	0.021	0.98	0.98	0.98
2	0.992	0.007	0.992	0.992	0.992

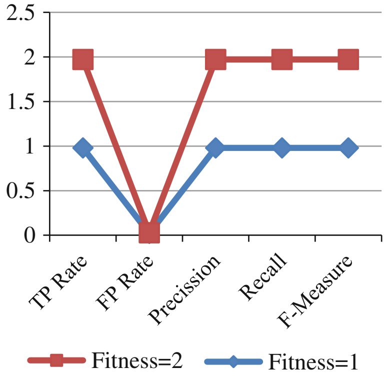


Fig. 5 Performance parameters for fitness

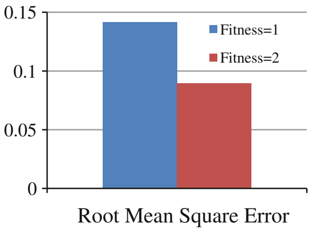


Fig. 6 Root mean square error for fitness values

5 Conclusion and Future Work

Anomaly detection is always a challenging task in research because of its dynamic nature. By adopting the data fusion approach, the feature vector fitness improved with this and the ensemble classifier improves the anomaly detection rate as well as detection speed. We tried the experimentation with HTTP dataset CSIC 2010; but in this, we have found few attributes, results compared with KDDCUP'99 dataset more feasible for NIDS.

In future, genetic algorithm and rough fuzzy techniques can be combined and applied to improve the data quality further; then it classifies the data with SVM for enhancing the performance and accuracy of real-time intrusion detection.

References

1. Lee W and Stolfo S., "Data Mining techniques for intrusion detection", In: Proc. of the 7th USENIX security symposium, San Antonio, TX, 1998.
2. L. Wenke, S. J. Stolfo, and K. W. Mok, " A data mining framework for building intrusion detection models:, in proc. IEEE Sump. Security Privacy, 1999, pp. 120–132.

3. Denning D. (1987) “An Intrusion-Detection Model,” IEEE Transactions on Software Engineering, Vol. SE-13, No. 2, pp. 222–232.
4. The Royal Road for Genetic Algorithms: Fitness Landscapes and GA Performance”, in: Francisco J. Varela, Paul Bourgine, editors. Toward a Practice of Autonomous Systems: proceedings of the first European conference on Artificial Life (1992). MIT Press.
5. Pawlak Z: Rough sets Present state and the future. Foundations of computing and Decision sciences 18, 157–163 (1993).
6. Pawlak Z: Rough Sets and Intelligent Data Analysis, Information Sciences, 2002, 147:1–12.
7. Pawlak Z, Rough Sets, International Journal of Computer and Information Sciences, vol. 11, pp. 341–256, 1982.
8. Boussouf M (1998) A Hybrid Approach to Feature Selection. Lecture Notes in Artificial Intelligence 1510:231–238.
9. Jan G. Bazan, Marcin Szczuka, “The rough set exploration system (2005)” TRANSACTIONS ON ROUGH SETS III, Springer.
10. http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html.
11. Tavallaee. M, Bagheri, E.; Wei Lu; Ghorbani, A.A., “A detailed analysis of the KDD CUP 99 data set,” Computational Intelligence for Security and Defence Applications, 2009. CISDA 2009. IEEE Symposium on, vol., no., pp. 1,6, 8–10 July 2009.
12. John R. Koza. Genetic Programming. MIT Press, 1992.
13. Man Leung Wong, Kwong Sak Leung, Data Mining Using Grammar based Genetic Programming and Applications, Kluwer Academic Publishers, 2000.
14. D.E.Goldberg, Genetic Algorithm in Search Optimization and Machine Learning. Reading, MA: Addison Wesley, 1989.
15. W. Lu and I. Traore, “Detecting new forms of network intrusion using genetic programming”, Computational Intelligence, Vol. 20, no 3, pp. 474–494, 2004.
16. H. Iba, “ Bagging, boosting and bloating in genetic programming”, in Proc. Genetic Evol. Comput. Conf., W. Banzhaf et al., Eds., 1999, pp. 1053–1060.

NUNI (New User and New Item) Problem for SRSs Using Content Aware Multimedia-Based Approach

Pankaj Chaudhary, Aaradhana A. Deshmukh, Albena Mihovska and Ramjee Prasad

Abstract Recommendation systems suggest items and users of interest based on preferences of items or users and item or user attributes. In social media-based services of dynamic content (such as news, blog, video, movies, books, etc.), recommender systems face the problem of discovering new items, new users, and both, a problem known as a cold start problem, i.e., the incapability to provide recommendation for new items, new users, or both, due to few rating factors available in the rating matrices. To this end, we present a biclustering technique, a novel cold start recommendation method that solves the problem of identifying the new items and new users, to alleviate the dimensionality of the item-user rating matrix using biclustering technique. To overcome the information exiguity and rating diversity, it uses the smoothing and fusion technique. As discussed, the system presents content aware multimedia-based social recommender media substance from item and user bunches.

Keywords Social recommender system · Cold start problem · Social media sites · Ratings · Biclustering · Rating matrix

P. Chaudhary (✉)
Smt. Kashibai Navale College of Engineering, Pune, India
e-mail: pankaj3253@gmail.com

A.A. Deshmukh · A. Mihovska · R. Prasad
Department of Electronic Systems, Center of TeleInfrastuktur (CTIF),
Aalborg University, Aalborg, Denmark
e-mail: aad@es.aau.dk

A. Mihovska
e-mail: albena@es.aau.dk

R. Prasad
e-mail: prasad@es.aau.dk

H.S. Saini et al. (eds.), *Innovations in Computer Science and Engineering*,
Advances in Intelligent Systems and Computing 413,
DOI 10.1007/978-981-10-0419-3_34

1 Introduction

The recommendation system has an important component in social media sites (such as Amazon, IMDB, and MovieLens) [1]. A key challenge in the recommender system is how to give recommendations to new users, new items, or both, a problem known as a cold start problem and also called new item and new user problem. It could fail the quality for the new users and new items, because the system knows very little about these items and users in terms of their preferences [2].

Thus, in this paper, we have presented a novel technique, which enables the social media substance to be repaired from the impact of the cold start problem in social recommender system. A novel technique to solve the new item and new user problem (cold start problem) is the biclustering technique. It can technically recognize the rating source for recommendations, and we have used the most plausible items and frequent raters. For reducing the dimensionality of the item-user rating matrix, we use the biclustering technique. To overcome the information exiguity and rating diversity, it uses the smoothing and fusion technique. Thus the system presents a content aware multimedia-based social recommender system.

2 Related Work

2.1 Biclustering

The biclustering technique (two-mode clustering) simultaneously clusters both item and user in an item-user matrix. The biclustering technique performs better than a one-way cluster technique to deal with sparse and high-dimensional recommendation matrices. Biclustering allows us to trust only a subset of attributes when looking for sameness between the objects. The aim of biclustering is to discover submatrices in the dataset, i.e., subsets of attributes and subsets of features [3, 4].

2.2 Collaborative Filtering with the Biclustering Technique

Collaborative filtering (CF) is a process that chooses items based on the correlation between people with similar preference from large-scale item-user matrices. It is also based on the idea that people who agreed in their evaluation of some items in the past are likely to agree again in the future [5]. A considerable number of the CF plots principally utilize one-dimensional clustering to cluster items or users exclusively. In any case, the one-dimensional clustering systems typically neglect the valuable data in the inverse dimension. The biclustering method, on the other hand, clusters both the item dimension and the user dimension in the item-user matrix [6–8]. In [9], Victor et al. proposed a trust-based CF scheme; this scheme incorporated the trust network through all the users in SRSs into the ratings. It achieved high level of recommendation accuracy by identifying and leveraging key features on the trust networks. However, they all assumed that the added external information was ready to be included. It invited some high cost to supplement the profiles of the various users and items. In [10], Sachin et al. proposed that the gaps between the indications of existing and newly arrived users or items are spontaneous. This allows the CF to handle the cold start problem.

In [11], George et al. proposed the CF approach key idea to simultaneously obtain user and item neighborhoods via co-clustering. This approach can provide high-quality predictions at a much lower computational cost, but the prediction of the ratings depends only on the summary statistics. In [12], Lee et al. proposed a personalized digital television (DTV) program recommendation system. This system refined the channels, selecting different processes for satisfying the requirements of the customer. TV contents recommender systems have many limitations caused from disadvantages of CF. In [13], Lai et al. proposed a system called cloud-based program recommendation system (CPRS). This system was used to recommend the programs to the customers of digital TV platforms and provides a scalable and powerful back-end to support large-scale data processing for a program recommender system, but network speed limits may become a significant problem when the public cloud is used.

3 Proposed System

3.1 Problem Statement

Most of the RSs suffer from the cold start problem. A key challenge in a recommender system is how to give recommendations to new users, new items, or both, which constitutes the cold start problem and is also known as the new item and new user problem. The cold start problem (new user and item problem) occurs in situations when social media site fails recommendation for new user and new item or both. This imagination is against the way that lesser or fewer ratings are given by users.

3.2 Model

We propose a content aware multimedia-based social recommender system. The recommender system fuses the browsing history of the users in a community/group and creates a liking profile for each participating user. It creates a liking profile by storing the metadata (title, rating, description) of the multimedia that a user listens

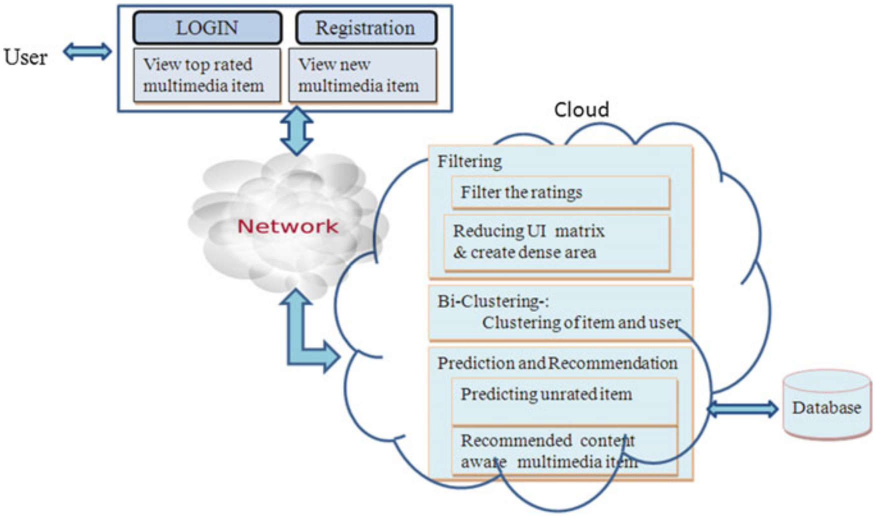

User
LOGIN
Registration
View top rated multimedia item
View new multimedia item
Network
Cloud
Filtering
Filter the ratings
Reducing UI matrix & create dense area
Bi-Clustering-: Clustering of item and user
Prediction and Recommendation
Predicting unrated item
Recommended content aware multimedia item
Database

Fig. 1 System architecture

to or watches. The system uses this liking profile to find other users and related multimedia content by matching the interests/content and uses that knowledge to recommend it (Fig. 1).

Proposed architecture follows using three parameters

(1) **Similarity**: The similarity of items and users can be measured by the Pearson correlation coefficient (PCC). PCC obtains better performance.

$$S_{x,y} = \frac{\sum_u \left(r_{u,x} - \overline{r_x}\right) * \left(r_{u,y} - \overline{r_y}\right)}{\sqrt{\sum_u \left(r_{u,x} - \overline{r_x}\right)^2} * \sqrt{\sum_u \left(r_{u,y} - \overline{r_y}\right)^2}} \tag{1}$$

where
$\overline{r_x}$ and $\overline{r_y}$ = the average ratings given to items x and y, respectively.
$r_{u,x}$ = rating given by user u on item x and $r_{u,y}$ = rating given by user u on item y.

(2) **Prediction**: The item-based CF predicts that an active user u likes the active item i

$$P_{u,x} = \overline{r_x} + \frac{\sum_{y \in I_u} S_{x,y} * \left(r_{u,y} - \overline{r_y}\right)}{\sum_{y \in I_u} S_{x,y}} \tag{2}$$

where

$P_{u,x}$ = predicted rating on the item x by the user u.

$\overline{r}_x$ and $\overline{r_y}$ = the average ratings given to items x and y, respectively.

$r_{u,\ y}$ = rating given by user u on item y and $S_{x,\ y}$ = similarity of items.

(3) **Accuracy**: Two general metrics which are used to evaluate the accuracy are as follows: mean absolute error (MAE) and the root mean squared error (RMSE).

$$\mathrm{MAE} = \frac{1}{|T|}\sum\nolimits_{u \in T} |P_{u,x} - r_{u,x}| \tag{3}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{|T|}\sum\nolimits_{u \in T} (P_{u,x} - r_{u,x})^2} \tag{4}$$

where

T = test set, $|T|$ = size of the test set, $P_{u,x}$ = predicted rating, and $r_{u,\ y}$ = actual rating.

3.3 Mathematical Model

Let S be a system such that

$$S = \{I, E, \mathrm{In}, \mathrm{Out}, T, f_{\mathrm{me}}, \mathrm{DD}, \mathrm{NDD}, \mathrm{MEM}_{\mathrm{shared}}, \mathrm{CPUCoreCnt}, \phi\}$$

where

S Proposed system and I = initial state at T < init > i.e., dataset

E End state of detecting multimedia recommendation

In is input of system and Out is output of system

T Set of serialized steps to be performed in pipelined machine cycle. In a given system, serialized steps are recommend item, view recommend item, etc.

f_{me} Main algorithm resulting into outcome Y

DD Deterministic data helps in identifying the assignment function. In a given system, deterministic data will recommend content aware video

NDD For non-deterministic data, in a given system we need to find the time required to recommend video

$\mathrm{MEM}_{\mathrm{shared}}$ Memory required to process all operations

CPUCoreCnt More the number of counts double the speed and performance

Φ Null value if any.

3.4 *Results and Discussion*

3.4.1 Dataset

We carry out a set of modern version of the popular Movie Lens dataset and so the similar file formats and metadata structures. Our dataset comprises two files: ratings.dat and movies.dat which, respectively, store the ratings and movie metadata. It contains over 50,000 ratings provided by more than 10,000 users on 5,000 unique items. Since the dataset is collected from social media, the expansion of the rated items (i.e., movies) is very large, leading to a sparsity value of at least 0.9993. This dataset is unfiltered.

3.4.2 Result

From the experimental approach, the obtained results are discussed. Clustering algorithm is divided into two categories: partition clustering and hierarchical clustering. This paper discusses one partition clustering algorithm (k-means) and hierarchical clustering algorithm. The figures show time comparison and testing accuracy between simple k-means and hierarchical clustering algorithm (Figs. 2, 3, and 4).

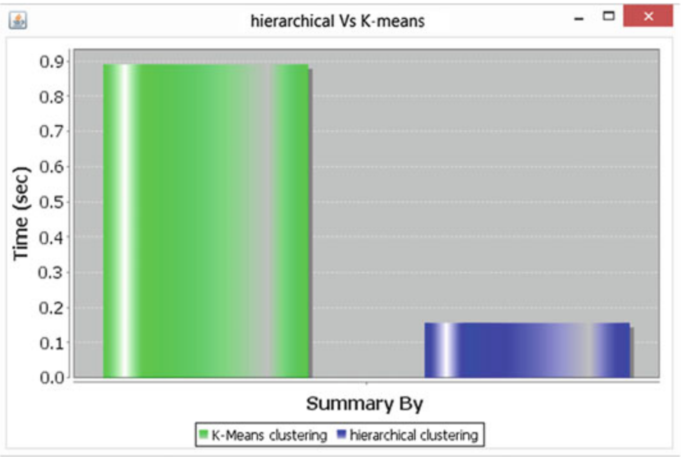


Fig. 2 Time comparison of k-means and hierarchical algorithm

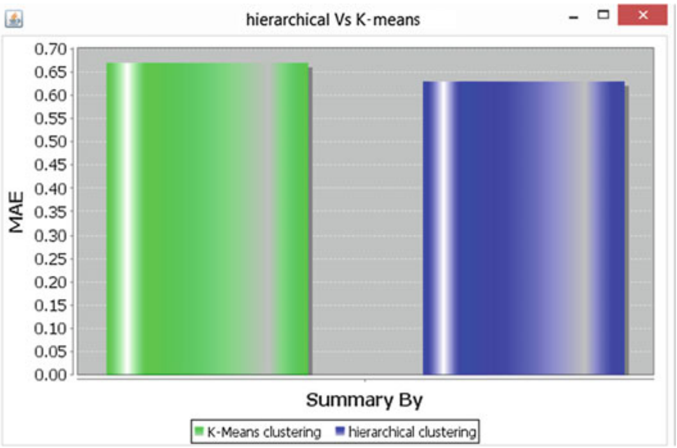


Fig. 3 MAE comparison of k-means and hierarchical algorithm

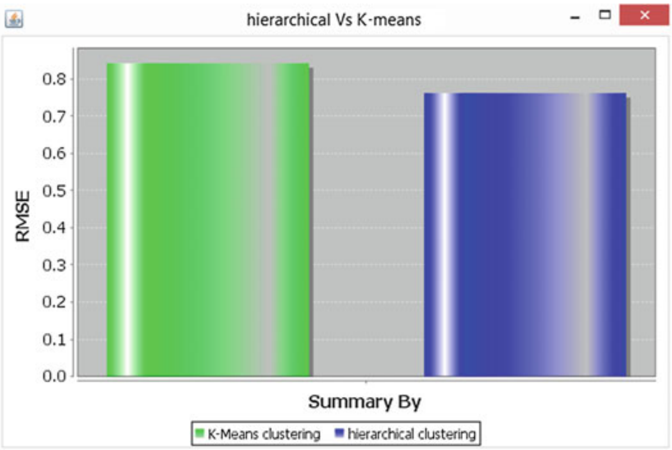


Fig. 4 RMSE comparison of k-means and hierarchical algorithm

4 Conclusions

The cold start problem is the core issue that should be addressed in the social recommender system. In this paper, we proposed a novel scheme, which improves the impact of the cold start problem on the recommender system. A novel scheme for solving the cold start problem is the biclustering technique. In the proposed technique, the rating sources for recommendation are distinguished and we used popular items and frequent raters. This would help in accurately identifying the similar items and the like-minded users. This system, however, could be further enhanced. We have developed it as a cloud service. We also plan to make it as a general social recommender container, to support different SRSs.

Further, we proposed a content aware multimedia-based social recommender system. The recommender system fuses the browsing history of users in a community/group to create a liking profile for each participating user by storing the metadata (title, rating, description, comments, view count, etc.) of the multimedia that a user listens to or watches.

References

1. F. Ricci, L. Rokach, B. Shapira, P. B. Kantor, “Recommender Systems Hand-book”, New York, NY, USA: Springer-Verlag, Oct. 2011.
2. D. Zhang, Q. Zou, and H. Xiong, “CRUC: Cold-start recommendations using collaborative filtering in internet of things”, Energy Proc., 2011.
3. Pankaj Chaudhary and A. A. Deshmukh, “A Survey of Content Aware Video based Social Recommendation System”, in Proc. IJSR, Volume 4 Issue 1, January 2015.
4. Hongya Zhao and Hong Yan et. al. “Biclustering Analysis for Pattern Discovery: Current Techniques, Comparative Studies and Applications”.
5. Xiaoyan Su and Taghi M. Khoshgoftaar,”A Survey of Collaborative Filtering Techniques”, in proc. of hindwai publication co., adv. In artifialinte. Volume 2009.
6. D. Zhang, J. Cao, M. Guo, J. Zhou, and R. Vaskar, “An efficient collaborative filtering approach using smoothing and fusing”, in Proc. ICPP, 2009.
7. K. W.-T. Leung and W.-C. Lee, “CLR: A collaborative location recommendation framework based on co-clustering”, in Proc. 34th Int. ACM SIGIR Conf., 2011.
8. Zhang et. al., “Cold Start Recommendation using Bi-Clustering and Fusion for large Scale Social Recommender Systems”, IEEE Transactions on Consumer Electronics, June 2014.
9. P. Victor, C. Cornelis, A. Teredesai, and M. De Cock, “Whom should I trust?: The impact of key figures on cold start recommendations”, in Proc. ACM Symp. Appl. Comput., 2008.
10. A. Schein, A. Popescul, L. Ungar, and D. Pennock, “Generative models for cold-start recommendations”, in Proc. SIGIR Workshop.
11. T. George and S. Merugu, “A scalable collaborative filtering framework based on co-clustering”, in Proc. 5th IEEE ICDM, Nov. 2005.
12. S. Lee, D. Lee, and S. Lee, “Personalized DTV program recommendation system under a cloud computing environment”, IEEE Trans. Consumer Electron., 2010.
13. C.-F. Lai, J.-H. Chang, C.-C. Hu, Y.-M. Huang, “CPRS: A cloud-based program recommendation system for digital TV platforms”, Future Generat. Comput. Syst., 2011.

Computerized Evaluation of Subjective Answers Using Hybrid Technique

Himani Mittal and M. Syamala Devi

Abstract To ensure quality in education, this paper proposes and implements a hybrid technique for computerized evaluation of subjective answers containing only text. The evaluation is performed using statistical techniques—Latent semantic analysis (LSA) and bilingual evaluation understudy (BLEU) along with soft computing technique, fuzzy logic. LSA is used to identify the semantic similarity between two concepts. BLEU helps prevent overrating a student answer. Fuzzy logic is used to map the outputs of LSA and BLEU. The hybrid technique is implemented using Java programming language, MatLab, Java open source libraries, and WordNet—a lexical database of English words. A database of 50 questions from different subjects of Computer Science along with answers is collected for testing purpose. The accuracy of the results varied from 0.72 to 0.99. The results are encouraging when compared with existing techniques. The tool can be used as preliminary step for evaluation.

Keywords Subjective evaluation · Latent semantic analysis · Bilingual evaluation understudy · Fuzzy logic · WordNet

1 Introduction

Evaluation is a systematic determination of a subject's merit, worth, and significance, using criteria governed by a set of standards. The primary purpose of evaluation is to gain insight into student learning and knowledge enhancement. Manual evaluation of subjective answers has limitations like time consuming, delayed result declaration, availability of experts, and scope for bias. There is a

H. Mittal (✉) · M. Syamala Devi
Department of Computer Science and Applications, Panjab University, Chandigarh, India
e-mail: research.himani@gmail.com

M. Syamala Devi
e-mail: syamala@pu.ac.in

H.S. Saini et al. (eds.), *Innovations in Computer Science and Engineering*,
Advances in Intelligent Systems and Computing 413,
DOI 10.1007/978-981-10-0419-3_35

need to develop intelligent software that can handle these problems efficiently and provide as much accuracy as possible compared to manual evaluation. If not replacing, it can aid the human examiner. In this paper, the evaluation of subjective answers is performed using a hybrid of statistical techniques and soft computing technique. Statistical technique for information retrieval, latent semantic analysis (LSA) [1], and technique for machine translation, bilingual evaluation understudy (BLEU) [2], are modified and used along with soft computing technique, fuzzy logic. There are several statistical techniques used for subjective evaluation, namely, latent semantic analysis, generalized latent semantic analysis, maximum entropy, and probabilistic latent semantic analysis. The latent semantic analysis was selected for this work because it is a clustering technique unlike all other techniques, which are classification techniques. The clustering technique is not dependent on any pre-specified classes/categories. It identifies the classes itself. The classification techniques need as input the categories and therefore cannot model the unknown. However, LSA has two drawbacks. First, LSA cannot distinguish between necessary and unnecessary repetition of keywords. To deal with this, BLEU is combined with LSA. BLEU acts as a clip on the maximum keyword usage. Second, LSA looks for exact word matches and does not consider the grammatical significance of the word in sentence structure. To overcome this problem lexicon ontology WordNet [3] is applied on the initial input. Soft computing technique and fuzzy logic are used to map the outputs of LSA and BLEU. The relationship between outputs of LSA and BLEU is defined using fuzzy logic.

The paper is organized as follows. The review of related work is given in Sect. 2. Section 3 discusses the methodology, tools, and techniques used for subjective evaluation. Section 4 includes implementation and testing details. Section 5 includes discussion of results and comparison with results of other techniques. Conclusions and scope for further work are presented in Sect. 6.

2 Review of Related Work

In 1994, Project Essay Grade (PEG) [4] was developed for automated English essay evaluation. It performs the evaluation based on features like essay length, word length, and vocabulary used. The reported accuracy of results is 83–87 %. However, it is argued that it does not take content into account.

In 1999, Foltz et al. [1] applied mathematical technique called latent semantic analysis (LSA) to computerized evaluation in a tool called intelligent essay assessor (IEA). In this method a matrix is made with keywords to be searched as rows and documents as columns. The frequency of each word in each document is recorded. Then singular value decomposition is done on this matrix. The reported correlation between LSA and human-assigned grades varied from 0.59 to 0.89. The correlation between two human graders is from 0.64 to 0.84. So the performance of LSA and human graders is comparable.

In 2005, Perez et al. [5] developed a system using latent semantic analysis (LSA) and BLEU (bilingual evaluation understudy) algorithm to essay evaluation. LSA performs semantic analysis and modified BLEU as used by the authors performs syntactic analysis. The results of the two are combined by a linear equation. However, the amount of weightage that should be given to BLEU-generated score and LSA-generated score is not fixed. Author has shown multiple combinations and average success rate is 50 %. In our paper, we have extended this work by combining the scores generated by LSA and BLEU using fuzzy logic.

Electronic essay rater (E-Rater) was developed by Attali and Burstein [6]. It used MSNLP (Microsoft Natural Language Processing) tool for parsing the text and extracting text features like word occurring probability, essay length, word length, vocabulary level used, and correlation with training essays. Then weightage is assigned to these features. Whenever a new essay is to be evaluated, its features are compared to already graded essays. It is successfully used for AWA (Analytical Writing Assessment) test in GMAT (Graduate Management Admission Test) with agreement rates between human expert and system consistently between 0.87 and 0.93. However, this tool is limited to evaluation of expression and grammar. Its applicability to evaluation of technical answers is unexplained.

In 2008, Kakkonen et al. [7] developed automatic essay assessor (AEA) that utilizes information retrieval techniques such as LSA, PLSA (probabilistic latent semantic analysis), and LDA (latent Dirichlet allocation) for automatic essay grading. Theoretically, PLSA and LDA are better models as they are generative models of LSA. LSA studies hidden values and clusters documents into different groups on the basis of relation and similarity between them. All the documents must be available when LSA is applied. It cannot predict any other variable or add new variables at any time. For adding new variables the analysis needs to be done again. It does not model the problem using variables. PLSA on the other hand uses probability to factorize the variables. It generates a model that can classify the documents. The classes are established as hidden variables. So it can create a model for all the already identified classes. For a new class the model cannot be applied directly. LDA is complete generative model which can help in modeling a new class also. They have tested the system with a set of 150 essays from an undergraduate course in education. LSA has an accuracy of 78 %, PLSA has 75 %, and LDA has 68 %. The experiments show that LSA has better performance.

In 2010, Islam and Hoque [8] proposed a system that makes use of generalized latent semantic analysis (GLSA) technique for evaluation. In GLSA instead of single term, word groups called n-gram are used for matrix construction. The reported accuracy of results is 89–96 %.

In 2010, Cutrone and Chang [9] in their research paper proposed a short answer evaluation method using natural language processing (NLP) techniques. This technique reduces the standard answer and student answer into its canonical form and compares them. Canonical forms of the standard and student answer were found using techniques like tokenization, stemming, morphological variation, etc. It can evaluate single-sentence answers only.

In 2011, Sukkarieh [10] discussed the max-entropy technique used in C-rater tool. Maximum entropy accepts as input categories database, i.e., database of pre-graded essays. It constructs an evaluation model from these categories by extracting features like what word precedes a given word. The student answers are also modeled in a similar manner. Neural network—perceptron—is trained using categories and the new inputs are tested. It has a reported accuracy of 0.48–0.98.

3 Methodology Used for Evaluation

The hybrid evaluation technique is implemented as a series of steps shown in Fig. 1. The inputs are all the student answers and one standard answer per question. The output is final marks of the students. First, preprocessing of input is done to prepare it for use in evaluation. The tokenization, synonym search, and stemming of student answers and standard answer are done. After the preprocessing, latent semantic analysis (LSA) and bilingual evaluation understudy (BLEU) techniques are applied independently. LSA measures the semantic relation between words using dimension reduction technique. BLEU calculates the word average and assigns marks. The output of LSA and BLEU are given as input to fuzzy logic. The output is generated as final marks of the student. The detailed working of the hybrid technique is given below.

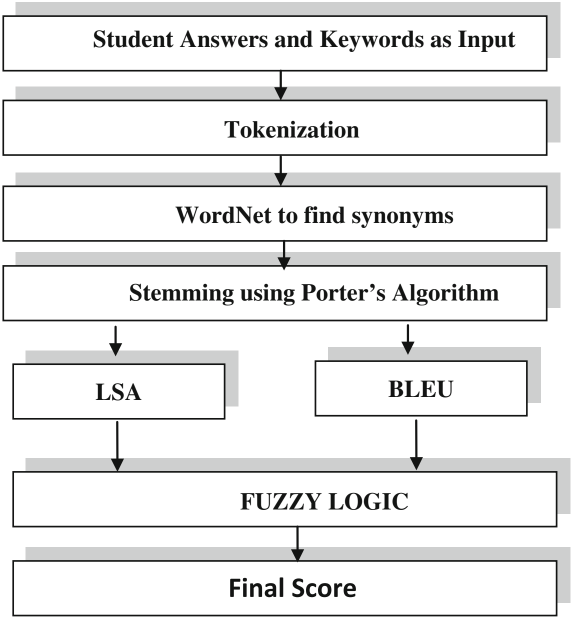

Student Answers and Keywords as Input
Tokenization
WordNet to find synonyms
Stemming using Porter's Algorithm
LSA
BLEU
FUZZY LOGIC
Final Score

Fig. 1 High-level steps in subjective evaluation

(1) Preprocessing steps are performed on the input answers and standard answer. First, tokenization is done to get individual words. Second, synonyms of all words in student answers and keywords are found. This allows for those answers in which student may not use standard words. Next, stemming is done to reduce the word to its basic stem using Porters stemming algorithm [11].

(2) Latent semantic analysis (LSA) [1] is a technique in natural language processing for analyzing relationships between a set of documents and the terms they contain. The basic assumption is that there exists a hidden semantic space in each text which is the accumulation of all words meaning. It usually takes three steps to compress the semantic space—filtering, selection, and feature extraction. The stop words are filtered. Then word frequency matrix is constructed by selecting reference texts. Then singular value decomposition is done to extract features by factorizing the feature matrix. LSA technique is applied on the preprocessed student answers. It generates the term frequency matrix (tdf) by keeping the terms as row heads and student answers as column heads. The term vectors and answer vectors are generated by performing singular valued decomposition of tdf matrix. These vectors represent individual terms and individual answers in the semantic 2-D plane. The cosine similarity of these vectors (correlation value) signifies the degree of relation between the student answer and the keywords. The semantic presence of the keywords in student answers is indicated by higher cosine (correlation) value.

(3) The BLEU [12] technique is used to overcome the drawback of LSA technique. LSA overrates the answers which repeat the keywords many times. BLEU generates a metric value ranging between 0 and 1. The value indicates how similar student answers are to the standard answer. The frequency of a keyword in student answer and total number of words in student answer are calculated. These values cannot be more than frequency of keyword in standard answer and total number of keywords in standard answer, respectively. If any value is more, then the corresponding value calculated from standard answer is used. It then divides frequency of each keyword in student answer and by the total number of words generating a fraction between 0 and 1.

4) The outputs of BLEU and LSA are mapped using fuzzy logic. Fuzzy logic is an extension of two-valued logic to handle the concept of partial truth. Compared to traditional crisp variables, a fuzzy variable has a truth value varying between 0 and 1 showing there degree of membership. Both LSA and BLEU give output as degree of correlation, so they are used as fuzzy input variables. These two independent variables are pointing toward the different aspects of level of similarity between standard answer and student answers. The values generated from LSA and BLEU for each student answer are given as input to the fuzzy logic to generate the final marks. The fuzzy logic model is designed with two input variables LSA and BLEU with three membership functions (bad, average, and excellent) and one output variable (Final) with four membership functions (bad, ok, average, and excellent). The technique uses Mamdani model. The rules are as follows:

If (LSA = bad) and (BLEU = bad) then result = bad;
If (LSA = bad) and (BLEU = average) then result = ok;
If (LSA = bad) and (BLEU = excellent) then result = ok;
If (LSA = average) and (BLEU = bad) then result = ok;
If (LSA = average) and (BLEU = average) then result = average;
If (LSA = average) and (BLEU = excellent) then result = average;
If (LSA = excellent) and (BLEU = bad) then result = ok;
If (LSA = excellent) and (BLEU = average) then result = average;
If (LSA = excellent) and (BLEU = excellent) then result = excellent.

4 Implementation and Testing

The hybrid technique is implemented using Java programming language, Java Agent Development Environment (JADE) [13], and several other tools like MatLab and WordNet. MatLab is used for performing singular value decomposition (SVD) for LSA and fuzzy logic design. WordNet [2] software is used for finding synonyms of the keywords. Open source libraries are used for invoking MatLab [14] and WordNet from java code [3].

4.1 Testing

There is no standard database in subjective answer evaluation which can be used for testing the hybrid technique. Therefore, the database was created by conducting class tests. The database consists of 50 questions with various answers (60–200 answers each) from field of Computer Science (technical answers). Questions are selected from subjects like computer organization and maintenance, C++, UNIX, Database Management, Project Management, Entrepreneur Development Program, Fundamentals of IT, Financial Management, eCommerce, and Operating Systems. The hybrid technique was applied to above database and evaluation was done. The same answers were given to human evaluator. The human-assigned scores and computer-generated scores are compared. The accuracy of results of hybrid technique varied between 0.72 and 0.99. A sample of the results generated is given in Table 1. The individual marks generated for sample answers of subject computer organization and maintenance are given along with human-assigned scores. The accuracy of hybrid technique is compared with the accuracy of existing tools, namely, E-rater, C-rater, Atenea, and intelligent essay assessor (IEA) and results are shown in Table 2. Table 2 also includes the number of questions used for testing and techniques used in development of each tool.

Table 1 Sample results generated using hybrid technique

Subject: computer organization and maintenance		
Question: explain the process of direct memory access		
Maximum marks: 5		
Student answers	Human-assigned marks	Hybrid technique-based software generated marks
The CPU transfer the system bus control to device manager. Then the data is transferred directly between device and memory. This facilitates bulk data transfer. For example, a sound card may need to access data stored in the computer's RAM, but since it can process the data itself, it may use DMA to bypass the CPU. Video cards that support DMA can also access the system memory and process graphics without needing the CPU. Ultra DMA hard drives use DMA to transfer data faster than previous hard drives that required the data to first be run through the CPU. In order for devices to use direct memory access, they must be assigned to a DMA channel. Each type of port on a computer has a set of DMA channels that can be assigned to each connected device. For example, a PCI controller and a hard drive controller each have their own set of DMA channels	4	3.52
Direct memory access is direct data transfer between memory and secondary storage device without CPU interference	1	1.41
I do not know what is direct memory access. IT is new term direct memory access is direct memory access and direct	0	0.1
The process where information is transferred between the primary memory and secondary memory is known as direct memory access. The CPU will temporarily release control on the system bus	2	2.04

Table 2 Comparison of results of hybrid technique-based evaluation with existing tools

Criteria \tool	IEA	E-rater	Atenea	C-rater	Hybrid technique
Accuracy maximum	89 %	93 %	79 %	98 %	99 %
Accuracy minimum	59 %	87 %	23 %	48 %	72 %
Number of questions	13	15	10	7	50
Technique (s) used	Latent semantic analysis	Latent semantic analysis, word average, and grammar-based feature extraction. Feature scores combined using linear regression	Latent semantic analysis and bilingual evaluation understudy combined using linear equation	Maximum entropy-based technique	Latent semantic analysis, bilingual evaluation understudy and fuzzy logic

5 Discussion of Results

The hybrid technique is capable of handling extreme cases of answers. Sometimes students, who do not know the answers, write invalid content like stories, songs, etc., and repetition of some keywords or question itself in the answer. Such extreme case answers are not given marks. Small and brief answers covering a number of valid keywords are given average marks.

The BLEU technique gives a low score if the exact match of keywords is less in the student answer. The score generated by BLEU can be considered as a lower bound on minimum marks to be awarded to the student. BLEU neither measures grammatical structure and expression nor performs semantic analysis. It is only checking exact word matches and dividing it with total number of keywords.

The LSA technique assigns score for the presence of keyword and semantic similarity of terms is also taken care. It does not consider the syntactic structure of the answers but measures the semantic aspect thoroughly. However, the scores generated using LSA technique can be used as an upper bound on the maximum marks that can be assigned to the student answer.

The hybrid technique combines the best features of LSA and BLEU. The syntactic structures of sentences and word similarity are taken care by the use of WordNet tool. The BLEU technique scores can be considered as lower bound and LSA technique scores can be considered as an upper bound on marks. This technique combines the two scores using fuzzy logic. This technique is able to identify invalid essays.

6 Conclusions and Scope for Future Work

The hybrid technique-based software can help to a large extent the human examiner in evaluating subjective answers. The techniques used for evaluation LSA and BLEU are complementary combination. The fuzzy function gives balanced weight to LSA and BLEU depending on different combinations of outputs. The use of WordNet helps in reduction of number of keywords to be given, as it finds synonyms of given keywords. This ensures student can make use of words of his choice. The performance of technique can be improved by introducing domain-specific ontology. The system can be enhanced to evaluate answers that include images, program code, equations, and other material.

References

1. Foltz, P. W., Laham, D., & Landauer, T. K. Automated Essay Scoring: Applications to Educational Technology. In B. Collis and R. Oliver (Eds.), Proceedings of EDMedia'99, Charlottesville, VA: Association of Computing in Education, (pp. 939–944), (1999).
2. WordNet—http://wordnet.princeton.edu/.

3. WordNet access Libraries—http://projects.csail.mit.edu/jwi/api/.
4. Page, E. B., Computer Grading of Student Prose, Using Modern Concepts and Software, The Journal of Experimental Education, 62(2), 127–142, (1994).
5. Daina Perez, Aflio Gliozzo, Carlo Strapparava, Automatic Assessment of students free-text answers underpinned by a combination of a BLEU-inspired algorithm and Latent Semantic Analysis, American Association for Artificial Intelligence, (2005).
6. Attali, Y. and Burstein, J., Automated Essay Scoring with e-rater V.2, The Journal of Technology, Learning and Assessment 4(3), (2006).
7. Kakkonen, T., Myller, N., Sutinen, E., Timonen, J., Comparison of Dimension Reduction Methods for Automated Essay Grading, Educational Technology & Society, 11(3), 275–288, (2008).
8. Islam M., Hoque A.S.M.L., Automated Essay Scoring Using Generalized Latent Semantic Analysis, Proceedings of 13th International Conference on Computer and Information Technology, pp. 358–363, (2010).
9. Cutrone, L., Chang, M., Automarking: Automatic Assessment of Open Questions, 10th IEEE International Conference on Advanced Learning Technologies, IEEE, 143–147, (2010).
10. Sukkarieh, J. Z., “Using a MaxEnt Classifier for the Automatic Content Scoring of Free-Text Responses”, American Institute of Physics Conference Proceedings, 1305(1), 41, (2011).
11. M.F. Porter, “An algorithm for suffix stripping”, Program: electronic library and information systems, Vol. 14 Iss: 3, pp. 130–137, (1980).
12. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W. J., “BLEU: a method for automatic evaluation of machine translation”. ACL-2002: 40th Annual meeting of the Association for Computational Linguistics. pp. 311–318, (2002).
13. JADE—www.jade.tilab.com.
14. Matlab Control—https://code.google.com/p/matlabcontrol/.

A Proposal: High-Throughput Robust Architecture for Log Analysis and Data Stream Mining

Adnan Rashid Hussain, Mohd Abdul Hameed and Sana Fatima

Abstract Various data mining approaches are now available, which help in handling large static data sets, in spite of limited computational resources. However, these approaches lack in mining high-speed endless streams, as their learning procedure though simple require the entire training process to be repeated for each new arriving information instance. The main challenges while dealing with continuous data streams: they are of sizes many times greater than the available memory, are real-time, and the new instances should be inspected at most once, and predictions must be made. Another issue with continuous real-time data is changing of concepts with time, which is often called concept drift. This paper addresses the above stated problems, and provides a solution by proposing a real-time, scalable, and robust architecture. It is a general-purpose architecture, based on online machine learning, which efficiently logs and mines the stream data in a fault-tolerant manner. It consists of two frameworks: (1) Event aggregation framework, which reliably collects events and messages from multiple sources and ships them to a destination for processing (2) Real-time computation framework, which processes streams online for extraction of information patterns. It guarantees reliable processing of billions of messages per second. Furthermore, it facilitates the evaluation of the stream learning algorithms and offers change detection strategies to detect concept drifts.

A.R. Hussain (✉)
Research & Development, Host Analytics Sofwtare Pvt. Ltd.,
Hyderabad 500 081, AP, India
e-mail: adnanrashid.ar@gmail.com

M.A. Hameed
Department of Computer Science, University College of Engineering,
Osmania University, Hyderabad, India
e-mail: hameed@gmail.com

S. Fatima
Department of Computer Science, M.J College of Engineering
and Technology, Hyderabad, India
e-mail: sana_maseeh@yahoo.com

H.S. Saini et al. (eds.), *Innovations in Computer Science and Engineering*,
Advances in Intelligent Systems and Computing 413,
DOI 10.1007/978-981-10-0419-3_36

Keywords Online · Throughput · Machine learning · Stream mining · Log analysis · Concept drift · Real-time · Robust

1 Introduction

A growing number of emerging business and scientific apps like satellite radar, stock market, transaction web log, real-time surveillance systems, telecommunication systems, sensor networks [1, 2], and other dynamic environments generate massive amounts of data. This continuously generated real-time, unbounded sequence of data called as a data stream [1–4]. In last decade, much research attention has been given to log processing and mining of data streams. It is demanding to mine streams as it helps in extraction of important knowledge, which is necessary to take crucial decisions in real-time. However, log analysis and extraction of information structures as models and patterns may pose many challenges such as storage, computational, and querying. Due to huge memory requirements and high storage costs it is nearly impossible to store the entire stream at once [3]. Traditional data mining techniques have fallen short in addressing the needs of data stream mining. These methods required the entire data to be first stored and then processed it using complex algorithms in several passes [1, 4].

To overcome the disadvantage of these techniques, many log aggregation systems were developed in the past, which processed the logs in single pass rather than processing in several passes. Few of the recently developed specialized distributed log aggregators include Yahoo's HedWig [2] Facebook's Scribe [5], and Yahoo's Data Highway [6]. These were mainly designed to collect and load the log data into a data warehouse or a persistent storage. However, they tend not to be good for log processing for few reasons. First, they are mainly built for offline consumers such as data warehousing applications that do periodic large loads rather than continuous consumption. Second, they are weak in distributed support. Finally, they do not consider throughput as their primary design constraint. The proposed architecture uses Apache's Kafka [7] and Cloudera's Flume [8] which not only ensures online data consumption but also is a reliable delivery of logs, which will be discussed in the later sections.

For mining of consumed logs, a number of big data analytics solutions like Berkley's Spark [9] have been built over the past few years. However, these systems focused on batch processing of large data sets. Although, there exist many real-time computation frameworks, such as Yahoo!'s S4 [10], which can deal with streaming data but such systems does not guarantee total fault tolerance. Thus, for real-time computation of logs, we rely on Storm [11], a highly distributed, fault-tolerant stream processing system, which has the ability to mine millions of tuples per second.

In this paper, we propose a real-time, stream processing system which aims at persistent storage of logs and mining of streams. Its components were selected after evaluating a dozen of best of breed technologies (discussed in Sects. 5 and 6). It has the ability to gather and process millions of messages per second reliably. The center ideas behind this architecture: (1) Efficiently collecting, aggregating, and moving large amounts of log data in reliable, fault-tolerant manner (2) Real-time computation of unbounded streams (3) Overcoming concept drift by using online machine learning techniques.

When the four best technologies will be integrated, the resulting system will be:

1. A high-throughput stream log processing system which can process hundreds of gigabytes of data and can accumulate billions of messages per day allowing the stream mining system to consume data at its own rate, and rewind the consumption whenever needed. Furthermore, it can publish messages at the rate of 50,000 and 400,000 (for message batches 1–50 respectively) and can consume 22,000 messages per second which is four times higher than the traditional systems.
2 Able to support reading of data from popular log stream types, such as Avro, Syslog, and Netcat. It can gather logs collected from hundreds of web servers, and automatically send those logs to a dozen of agents that write to persistent storage cluster. A complex event processing system which, for example, can be used to identify meaningful events from a flood of events, and then take actions on those events in real-time (for example, using it, one can extract emerging trends from the social networking sites and maintains them at the local and national level). It can do a continuous query and stream the results to clients in real-time. It can process a stream of new data and update databases in real-time. In case of a miss, it can replay that missed data (tuple) and process it again thereby ensuring guaranteed message processing and high fault tolerance. Thus, it can handle data velocities of tens of thousands of messages every second.
3 Able to solve the problem of concept drift, and handle very high magnitudes of data. For a period of 10 h the number of examples that can be handled by our architecture at full speed range from around 300 to 19,000 million, which is approximately 0.25–1.75 terabytes of data. Handling data volumes of this magnitude indeed offers a much cheaper solution rather than finding resources to store and retrieve several terabytes of data. When such high amount of data can be handled it becomes much easier to detect changing concepts from continuously arriving data streams.

Rest of the paper is structured as follows: Sect. 3 provides a general view of proposed architecture for log aggregation and data stream mining and the requirements of the system are discussed. Sections 4 and 5 describe the problems with early systems, and the frameworks of new system are discussed in detail. Section 6 discusses the online machine learning algorithms, the problem of concept drift and its solution. Section 7 concludes this paper.

2 Proposed Architecture

2.1 Overview

In this section, we formally introduce the architecture which solves the problems of existing log aggregation and stream mining systems. Its major concern is robustness. It is highly adaptable to memory, time constraints, and data stream rate. The system can be easily distributed over a cluster of machines, and new streams can be added on the fly. The open-source implementations include, Apache Kafka, Cloudera's Flume, Storm from Twitter and massive online analysis software environment.

The architecture consists of two parts as shown in Fig. 1. The first part, event aggregation framework, deals with logging and handling of data and its persistent storage. The second part, real-time computation framework, mines the data, provided by the event aggregation framework, by extracting important information and storing the predicted values. The working of architecture is as follows:

1. Many terabytes of data arrive from various sources which is reliably collected logs to a location where it will be further analyzed.
2. The logged data is distributed on multiple queues randomly (will be discussed in detail in Sect. 5) and is processed for extraction of information structures. Processing of multiple input streams is done to produce new streams.
3. After mining of streams, the results are passed to stream learning algorithms, which update their model and detect changing concepts.

3 Objectives of Architecture

For efficient handling of big data, there is a need for a high-throughput, reliable system which can operate in near real-time without compromising the scalability factor. The system must be distributed in all its tiers and must have the ability to deal with massive data streams from various sources. With this clearer view of the problem, here are what we see as the objectives of our system:

1. *Performance*: To handle bursty and potentially unpredictable data streams, the system must have the ability to process very high throughputs of messages with very low latency.
2. *Reliability*: Faults can happen at many levels like software applications can fail, machines can fail, excessive network congestion can happen, or a node may go down for maintenance. Thus, the architecture must be flexible enough to make sure that events make it to permanent storage.

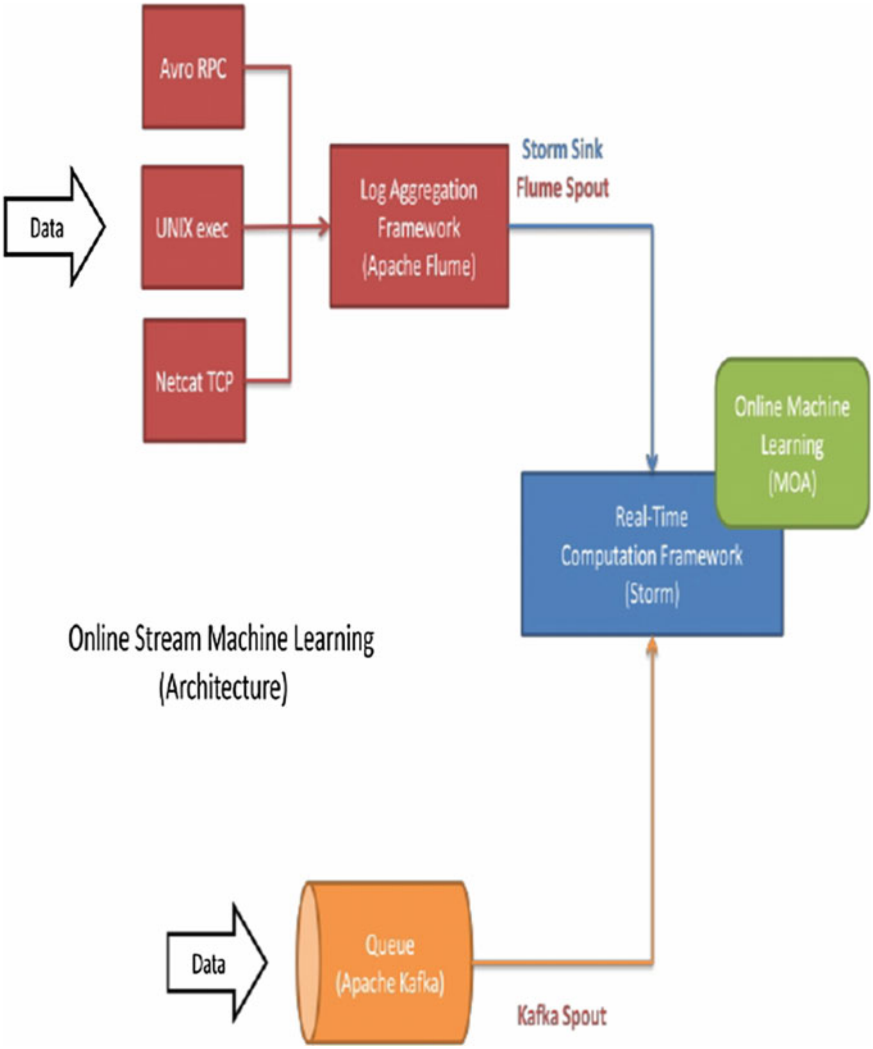

Avro RPC
Data
UNIX exec
Netcat TCP
Log Aggregation Framework (Apache Flume)
Storm Sink
Flume Spout
Online Machine Learning (MOA)
Real-Time Computation Framework (Storm)
Online Stream Machine Learning (Architecture)
Data
Queue (Apache Kafka)
Kafka Spout

Fig. 1 A high-throughput architecture for data stream mining

3. *Horizontal Scaling*: The system must have the capability to distribute processing across multiple processors and machines to achieve incremental scalability. Ideally, the distribution should be automatic and transparent.
4. *Real-time Computation*: The stream processing system must process time-series messages in a predictable manner to ensure that the results of processing are deterministic and repeatable.

5. *Robust*: The system must be susceptible to failures and should provide advanced recovery options in order not to lose data. It must also be durable enough to survive crashes and reboots.
6. *Learning*: The goodness of an online learning algorithm can be identified with the help of tools and methods used for evaluation. The architecture must provide realistically changing environment for testing and comparing various algorithms.

4 Event Aggregation Framework

The main job of this framework is to reliably collect messages and events from different sources, and forward it to the real-time computation framework (discussed in Sect. 5) without any loss of even a single event or message. For this purpose, the framework requires two components: (1) A highly distributed and fault-tolerant messaging system (2) A reliable log aggregator. Now, these components should make log collection as easy and reliable as possible, so that the streams from multiple sources can be efficiently captured without any loss. After evaluating a dozen of best of breed technologies drawn from the domains of distributed log collection, CEP/stream processing, and real-time messaging systems the components of this framework were selected.

4.1 Early Considerations, Their Drawbacks, and the New Technologies

1. *Messaging-Based System*: Earlier queues were considered for storage and reliable delivery of data streams. Thus, some of the existing queues were deeply studied, and we decided to choose from among the most commonly known and publicly available queues: Rabbit MQ [10] and ActiveMQ [9] for the purpose of event aggregation. But after evaluation based on parameters like reliability, failover rate, performance, latency, host discovery, etc., it was found that these queues lacked in one or the other parameter. The above-made observations show that neither of the queue has the ability to fully meet our objectives. For this purpose, we opted for Apache's Kafka, a highly distributed messaging-based system, which ensures the highest throughput and message persistence when compared to both the queues. Furthermore, an experimental study conducted by Jay Kreps et al. became the deciding factor in selection of Kafka. In this study, comparison of Kafka was done with both highly superior messaging systems, active MQ and Rabbit MQ, which showed that Kafka achieved a much higher throughput than both the systems. Also, it provided integrated distributed

support and showed the ability to scale out. The whole job of Kafka is to act as a shock absorber between the flood of events, and those who want to consume offline and online. In addition to it, to avoid log corruption, Kafka stores a CRC for each message in the log. It facilitates guaranteed message delivery by allowing the consuming application to "rewind" the consumption whenever needed and consume data at its own rate.

2. *Log Aggregator*: In order to meet the requirements of this component, few of the specialized log aggregators, namely Yahoo's HedWig [2], Facebook's Scribe [3], and Yahoo's Data Highway [1] were evaluated, and it was realized that all these log aggregators were mainly designed for offline log storage. If at all any aggregator supported online consumption, its reliability, throughput, and distributed support were not too strong to maintain continuous consumption of logs. Later, while evaluating these log aggregators, we came across Flume—an open-source distributed log aggregation software developed by Cloudera. It mainly is concerned with "online" gathering set of log files from a cluster of machines and aggregating them to a centralized persistent storage. It meets the four key goals: Reliability, Scalability, Manageability, and Extensibility. It offers a more integrated distributed support than the above-mentioned log aggregators.

5 Real-Time Computation Framework

We wanted to build an infrastructure which can support real-time computation of messages and events collected by the event aggregation framework. Thus, there was a requirement for a fault–tolerant, real-time computation system, without whose incorporation this framework would remain incomplete. Traditional big data analytics systems focused on batch processing and were suitable for dealing with offline data. However, when it comes to dealing with unterminated streams of data, we found Twitter's Storm; an open-source distributed complex event processing (CEP) system, to be one of the best solutions for distributed real-time processing. Although other big data solutions like Yahoo!'s S4, a high-performance computing (HPC) platform, exist we opted for Storm as they incorporate partial fault tolerance. What makes Storm most interesting is its easy integrability with both Flume and Kafka.

A typical Storm topology consists of spouts and bolts.

Streaming data arrives from external sources to spouts, which transfer the streams to bolts. Bolts perform transformation on streams originated by spouts by implementing various traditional and some complex functions like filtering, aggregations, or communication with external entities such as a database, etc. To support multiple transformations Storm topologies consist of a number of spouts and bolts [12, 13]. Storm ensures guaranteed message processing such that if a tuple emitted by spout remains unprocessed, then Storm replays the tuple from the spout until it is processed. Thus, Storm has got great ability to track unprocessed tuples.

An interesting feature of Storm is it supports detection of failures at task level, upon detecting a failure; messages are reassigned by quickly restarting the processing.

5.1 Integration of Event Aggregation Framework with Real-Time Computation Framework

Now that the general working of real-time computation framework is clearly understood, let us see how it deals with the messages coming from the event aggregation framework. After Storm's tight integration with Kafka and Flume, the events collected by them are passed to spouts. Each spout emits tuples which can be processed by one or more bolts. Bolts operate on tuples by performing many transformation functions, which were discussed above. As all the three systems are highly fault-tolerant and reliable, there is no chance that even a single message will be lost or remain unprocessed. With this fabulous integration, stream processing can be conducted at linear scale, assuring that every message gets processed in real-time, reliably. In tandem, both the frameworks can handle data velocities of tens of thousands of messages every second. This combination can form the best enterprise-grade real-time ETL and streaming analytics solution.

6 Online Machine Learning Algorithms

After real-time computation of messages, the data is now ready to be fed to the online machine learning algorithms. But before that, we first need to integrate real-time computation framework with an environment in which these algorithms can run, and their learning ability can be known. Therefore, MOA is opted as the final component which not only guarantees good comparison of various stream learning algorithms but also detects concept drifts.

Massive online analysis (MOA) is related to Weka, the Waikato environment for knowledge analysis, specific for mining data streams with concept drift. This software allows evaluation and comparison of algorithms and running experiments by providing the user with various tools and methods. It is mainly designed for analysis and comparison of stream mining algorithms within some explicit memory constraints [14]. Stream learning algorithms constantly update the learned model with continuously arriving examples without exceeding the time and memory limits. The output of these algorithms is a model completely updated with all the instances. To check the capability of algorithms, this framework provides large stream data generators and evaluation measures to compare the real-time scenarios and find the best fitting solutions.

In this section we discuss the problem of concept drift and its solution with MOA. Later, an explanation of MOA's integration with real-time computation framework will be given.

6.1 Concept Drift

Concept drift refers to changing of underlying distribution of data with time. Typical examples of concept drift include weather prediction rules and customer preferences. If sales are to be predicted for an online shop, the predictive model trained with earlier data becomes less accurate as time passes [15]. Thus, changing concepts often make the model built on new data inconsistent with old one. Therefore, much attention must be given to detect concept drift as predictions might become less accurate as time passes. Many approaches, existing in our literature, try to overcome this problem by proposing various algorithms, which help the learning models to adapt to changes accurately [16–18]. Therefore, to test the goodness of these algorithms, and their ability to identify true drifts we are using MOA framework. As MOA can handle data volumes of a very high magnitude, it becomes much easier to detect changing concepts from continuously arriving data streams.

6.2 Working of Real-Time Computation Framework with Online Machine Learning Algorithms

Finally, the real-time computation system is now integrated with the MOA framework. From the previous sections it's clear that the incoming streams from event aggregation framework are processed by bolts, and the output streams produced by bolts are then forwarded to MOA for evaluation. The stream mining algorithms running in this framework take the streaming data, and update the learning model constantly. The updated model then can predict new values for many real-time scenarios, for example, trending topics. The biggest advantage of using MOA is its ability to detect the changing concepts. As soon as concept drift is detected, the mining algorithm is notified about the change and it updates itself according to the new concept. So, apart from dealing big data efficiently, this architecture has got the ability to get informed about the changing concepts.

7 Conclusion

In this paper, we present a highly scalable, real-time, high-throughput, distributed architecture for log analysis and stream mining. It handles massively evolving data streams by first collecting and aggregating the log data from various generation [19]

Facebook's Scribe, points and then reliably storing and forwarding it for real-time computation. It provides two different frameworks for both log collection and mining of streams. Changing concepts are modeled and the learning models are constantly updated to overcome drifts. The fabulous combination of the two frameworks makes it highly fault-tolerant such that no single incoming message will be lost or remain unprocessed. Most importantly, the architecture is able to consume, publish, and process billions of events per second. The use of all open-source software's as its components makes it a true cost effective architecture.

References

1. Golab and Ozsu M. T.: Issues in Data Stream Management. In SIGMOD Record, Volume 32, Number 2, June (2003) 5–14.
2. Garofalakis M., Gehrke J., Rastogi R.: Querying and mining data streams: you only get one look a tutorial. SIGMOD Conference 2002: 35. (2002).
3. Babcock B., Babu S., Datar M., Motwani R., and Widom J.:Models and issues in data stream systems. In Proceedings of PODS (2002).
4. Muthukrishnan S.: Data streams: algorithms and applications. Proceedings of the fourteenth annual ACMSIAM symposium on discrete algorithms (2003).
5. http://developer.yahoo.com/blogs/hadoop/posts/2010/06/enabling_hadoop_batch_processi_1/.
6. https://issues.apache.org/jira/browse/ZOOKEEPER-775.
7. Kafka, http://sna-projects.com/kafka/.
8. Cloudera's Flume, https://github.com/cloudera/flume.
9. http://www.ibm.com/developerworks/library/os-spark/.
10. http://incubator.apache.org/s4/.
11. http://cloud.berkeley.edu/data/storm-berkeley.pdf.
12. Mohamed Medhat Gaber, Arkady Zaslavsky and Shonali Krishnaswamy. "Mining Data Streams: A Review", VIC3145, Australia, ACM SIGMOD Record Vol. 34, No. 2; June 2005.
13. http://activemq.apache.org.
14. Albert Bifet and Richard Kirkby. Massive Online Analysis, August 2009.
15. Alexey Tsymbal. (2004) The Problem of Concept Drift: Definitions and Related Work.
16. Bifet, A., Holmes, G., Pfahringer, B., Kirkby, R., Gavalda, R. (2009). New ensemble methods for evolving data streams. In 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.
17. Bifet, A. (2010). Adaptive Stream Mining: Pattern Learning and Mining from Evolving DataStreams, IOS Press.
18. Bifet, A. and Gavalda, R. (2007). Learning from Time-Changing Data with Adaptive Windowing, in SIAM Int. Conf. on Data Mining (SDM'07).
19. http://www.facebook.com/note.php?note_id=32008268919.

Author Index

H.S. Saini et al. (eds.), *Innovations in Computer Science and Engineering*,
Advances in Intelligent Systems and Computing 413,
DOI 10.1007/978-981-10-0419-3