Advances in Intelligent Systems and Computing 719

Pankaj Kumar Sa
Manmath Narayan Sahoo
M. Murugappan
Yulei Wu
Banshidhar Majhi *Editors*

Progress in Intelligent Computing Techniques: Theory, Practice, and Applications

Proceedings of ICACNI 2016, Volume 2

Springer

Advances in Intelligent Systems and Computing

Volume 719

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

Pankaj Kumar Sa · Manmath Narayan Sahoo
M. Murugappan · Yulei Wu
Banshidhar Majhi
Editors

Progress in Intelligent Computing Techniques: Theory, Practice, and Applications

Proceedings of ICACNI 2016, Volume 2

Springer

Editors
Pankaj Kumar Sa
Department of Computer Science
and Engineering
National Institute of Technology
Rourkela, Odisha
India

Manmath Narayan Sahoo
Department of Computer Science
and Engineering
National Institute of Technology
Rourkela, Odisha
India

M. Murugappan
School of Mechatronic Engineering
Universiti Malaysia Perlis (UniMAP)
Arau, Perlis
Malaysia

Yulei Wu
The University of Exeter
Exeter, Devon
UK

Banshidhar Majhi
Department of Computer Science
and Engineering
National Institute of Technology
Rourkela, Odisha
India

ISSN 2194-5357 ISSN 2194-5365 (electronic)
Advances in Intelligent Systems and Computing
ISBN 978-981-10-3375-9 ISBN 978-981-10-3376-6 (eBook)
DOI 10.1007/978-981-10-3376-6

Library of Congress Control Number: 2016963607

Printed on acid-free paper

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

Foreword

Message from the Honorary General Chair Prof. Mike Hinchey

Welcome to the 4th International Conference on Advanced Computing, Networking and Informatics. The conference is hosted at the Centre for Computer Vision and Pattern Recognition, NIT Rourkela, Odisha, India. For this fourth event, held during 22–24 September 2016, the theme is Computer Vision and Pattern Recognition.

Following the great success for the last 3 years, we are very glad to realize the event co-organized by the Center for Computer Vision and Pattern Recognition, at National Institute of Technology Rourkela, India; the Faculty of Engineering and Technology, Liverpool John Moores University, UK; the College of Engineering, Mathematics and Physical Sciences, University of Exeter, UK and the Faculty of Science, Liverpool Hope University, UI.

Having selected 114 articles from more than 500 submissions, we are glad to have the proceedings of the conference published in the *Advances in Intelligent Systems and Computing* series of Springer.

I am very pleased to have published special issues of papers from ICACNI in *Innovations in Systems and Software Engineering*: *A NASA Journal*, published by Springer and of which I am Editor-in-Chief, in each of the preceding years, and all of which were truly excellent and well received by our subscribers.

The papers accepted for this year will be considered again for this journal and for several other special issues.

I would like to acknowledge the special contribution of Prof. Sunil Kumar Sarangi, Former Director of NIT Rourkela, as the Chief Patron for this conference.

The conference is technically co-sponsored by the following professional organizations/laboratories:

1. Joint Control Systems Society and Instrumentation & Measurement Society Chapter, IEEE Kolkata Section
2. IEEE Communications Society Calcutta Chapter

3. Aerospace Electronics and Systems Division, CSIR National Aerospace Laboratories, Govt. of India
4. Dependable Computing and Networking Laboratory, Iowa State University, USA
5. Multimedia Content Security Innovative Research Group, Henan University of Science and Technology, China
6. Poznan University of Technology Vision Laboratory, Poland

We are grateful to all of them for their co-sponsorship and support. The diversity of countries involved indicates the broad support that ICACNI 2016 has received. A number of important awards will be distributed at this year's event including Best Paper Awards from ACM Kolkata Chapter, Best Student Paper Award from IEEE ComSoc Koltaka Chapter, Student Travel award from INNS, and Distinguished Women Researcher Award.

I would like to thank all of the authors, contributors, reviewers, and PC members for their hard work. I would especially like to thank our esteemed keynote speakers and tutorial presenters. They are all highly accomplished researchers and practitioners and we are very grateful for their time and participation.

However, the success of this event is truly down to the local organizers, local supporters, and various Chairs who have done so much work to make this a great event. We hope you will gain much from ICACNI 2016 and will plan to submit to and participate in ICACNI 2017.

Prof. Mike Hinchey
ICACNI 2016 Honorary General Chair
President, International Federation for Information Processing (www.ifip.org)
Director, Lero-the Irish Software Engineering Research Centre (www.lero.ie)
Vice Chair and Chair Elect, IEEE UK and Ireland Section
mike.hinchey@lero.ie

Preface

It is indeed a pleasure to receive overwhelming response from academicians and researchers of premier institutes and organizations of the country and abroad for participating in the 4th International Conference on Advanced Computing, Networking, and Informatics (ICACNI 2016), which makes our endeavor successful. The conference organized by Centre for Computer Vision and Pattern Recognition, National Institute of Technology Rourkela, India during 22–24 September 2016 certainly marks a success toward bringing researchers, academicians, and practitioners in the same platform. We have received more than 550 articles and very stringently have selected through peer review 114 best articles for presentation and publication. We could not accommodate many promising works as we tried to ensure the highest quality. We are thankful to have the advice of dedicated academicians and experts from industry and the eminent academicians involved in providing technical co-sponsorship to organize the conference in good shape. We thank all people participating and submitting their works and having continued interest in our conference for the fourth year. The articles presented in the two volumes of the proceedings discuss the cutting-edge technologies and recent advances in the domain of the conference. The extended versions of selected works would be re-reviewed for publication in reputed journals.

We conclude with our heartiest thanks to everyone associated with the conference and seeking their support to organize the 5th ICACNI 2017 at National Institute of Technology Goa during 01–03 June 2017.

Rourkela, India	Pankaj Kumar Sa
Rourkela, India	Manmath Narayan Sahoo
Arau, Malaysia	M. Murugappan
Exeter, UK	Yulei Wu
Rourkela, India	Banshidhar Majhi

The original version of the book was revised: Volume number of the book has been updated. The erratum to the book is available at https://doi.org/10.1007/978-981-10-3376-6_62

Advisory Board Members

Arun Somani, FIEEE, Iowa State University, USA
Costas StasoPoulos, IEEE Region 8 Director 2015–2016, Electricity Authority of Cyprus, Cyprus
Dacheng Tao, FIEEE, FIAPR, FIET, University of Technology, Sydney
Friedhelm Schwenker, University of Ulm, Germany
Ishwar K. Sethi, FIEEE, Oakland University, USA
Kenji Suzuki, SMIEEE, Illinois Institute of Technology, USA
Mohammad S. Obaidat, FIEEE, Fordham University, USA
Nikhil R. Pal, FIEEE, Vice President for Publications IEEE Computational Intelligence Society (2015–16), Indian Statistical Institute Kolkata, India
Rajkumar Buyya, FIEEE, The University of Melbourne, Australia
San Murugesan, SMIEEE, Western Sydney University Sydney, Australia
Sanjeevikumar Padmanaban, SMIEEE, Ohm Technologies, India
Subhas Mukhopadhyay, FIET, FIEEE, Massey University, New Zealand
Subhash Saini, The National Aeronautics and Space Administration (NASA), USA
Vincenzo Piuri, FIEEE, Vice President for Technical Activities IEEE 2015, University of Milano, Italy

Chief Patron

Sunil Ku. Sarangi, FNAE
Former Director
National Institute of Technology Rourkela, India

Patron

Banshidhar Majhi
Dean (Academic)

Professor, Department of Computer Science & Engineering
National Institute of Technology Rourkela, India

Honorary General Chair

Mike Hinchey, FIET, SMIEEE
President, International Federation for Information Processing
Director, Lero-the Irish Software Engineering Research Centre
Vice Chair and Chair Elect, IEEE UK and Ireland Section
Former Director and Expert, Software Engineering Laboratory,
NASA Goddard Space Flight Centre
Professor, University of Limerick, Ireland

General Chairs

Durga Prasad Mohapatra, National Institute of Technology Rourkela, India
Manmath Narayan Sahoo, National Institute of Technology Rourkela, India

Organizing Co-chairs

Pankaj Kumar Sa, National Institute of Technology Rourkela, India
Sambit Bakshi, National Institute of Technology Rourkela, India

Programme Co-chairs

Atulya K. Nagar, Liverpool Hope University, UK
Dhiya Al-Jumeily, Liverpool John Moores University, UK
Yulei Wu, The University of Exeter, UK

Technical Programme Committee

Abir Hussain, Liverpool John Moores University, UK
Adam Schmidt, Poznan University of Technology, Poland
Akbar Sheikh Akbari, Leeds Beckett University, UK
Al-Sakib Khan Pathan, SMIEEE, UAP and SEU, Bangladesh/Islamic University in Madinah, KSA
Andrey V. Savchenko, National Research University Higher School of Economics, Russia

Organizing Committee

Contents

Part I Cloud Computing, Distributed Systems, Social Networks, and Applications

Part II Applications of Informatics

Part III Authentication Methods, Cryptography and Security Analysis

Part IV Big Data and Recommendation Systems

Part V Emerging Techniques in Computing

Part VI Internet, Web Technology, Web Security, and IoT

Part VII Research on Optical Networks, Wireless Sensor Networks, VANETs, and MANETs

Part VIII Communication Systems, Antenna Research, and Cognitive Radio

About the Editors

Dr. Pankaj Kumar Sa received Ph.D. degree in Computer Science in 2010. He is currently serving as Assistant Professor with the Department of Computer Science and Engineering, National Institute of Technology Rourkela, India. His research interests include computer vision, biometrics, visual surveillance, and robotic perception. He has co-authored a number of research articles in various journals, conferences, and book chapters. He has co-investigated some research and development projects that are funded by SERB, DRDOPXE, DeitY, and ISRO. He has received several prestigious awards and honors for his excellence in academics and research. Apart from research and teaching, he conceptualizes and engineers the process of institutional automation.

Dr. Manmath Narayan Sahoo is Assistant Professor in Computer Science and Engineering Department at National Institute of Technology Rourkela, Rourkela, India. His research interests include fault-tolerant systems, operating systems, distributed computing, and networking. He is the member of IEEE, Computer Society of India, and The Institutions of Engineers, India. He has published several papers in national and international journals.

Dr. M. Murugappan is Senior Lecturer in School of Mechatronic Engineering at Universiti Malaysia Perlis (UniMAP), Perlis, Malaysia. He received his Ph.D. degree in Mechatronic Engineering from Universiti Malaysia Perlis (UniMAP), Malaysia in 2010, Master of Engineering degree in Applied Electronics from Government College of Technology, Anna University, Tamilnadu, India in 2006 and Bachelor of Electrical & Electronics Engineering from Adiparasakthi Engineering College, Melmaruvathur, Tamilnadu in 2002. His research interest include signal processing (EEG, ECG, HRV, ECG), affective computing (emotion, stress, emotional stress), pattern recognition, brain computer interface (BCI), human machine interaction (HMI), digital image processing, statistical analysis, neuromarketing and neurobehavioral analysis. He has published over 45 research papers in refereed journals and over 50 papers in national and international conferences.

Dr. Yulei Wu is Lecturer in Computer Science at the University of Exeter. He received his Ph.D. degree in Computing and Mathematics and B.Sc. degree in Computer Science from the University of Bradford, UK, in 2010 and 2006, respectively. His recent research focuses on future network architecture and protocols, wireless networks and mobile computing, cloud computing, and performance modeling and analysis. He has published over 30 research papers on these areas in prestigious international journals, including IEEE Transactions on Parallel and Distributed Systems, IEEE Transactions on Communications, IEEE Transactions on Wireless Communications, IEEE Transactions on Vehicular Technology and ACM Transactions on Embedded Computing Systems, and reputable international conferences. He was the recipient of the Best Paper Awards from IEEE CSE 2009 and ICAC 2008 conferences. He has served as the Guest Editor for many international journals including Elsevier Computer Networks and ACM/Springer Mobile Networks and Applications (MONET). He has been the Chair or Vice-Chair of 20 international conferences/workshops and has served as the PC member of more than 60 professional conferences/workshops. He was awarded the Outstanding Leadership Awards from IEEE ISPA 2013 and TrustCom 2012, and the Outstanding Service Awards from IEEE HPCC 2012, CIT 2010, ScalCom 2010. His areas of interest are future Internet architecture that include software-defined networking, network functions virtualization, clean-slate post-IP network technologies (e.g., information centric networking), cloud computing technologies, and mobility; wireless networks and mobile computing; cloud computing; and performance modeling and analysis.

Dr. Banshidhar Majhi is Professor in Computer Science & Engineering Department at National Institute of Technology Rourkela, Rourkela, India. Dr. Majhi has 24 years of teaching and 3 years of industry experience. He has supervised 8 Ph.D. students, 40 M.Tech, 70 B.Tech students, and several others are pursuing their courses under his guidance. He has over 50 publications in journals and 70 publications in conference proceedings of national and international repute. He was awarded with Gold Medal for Best Engineering Paper from IETE, 2001 and from Orissa Engineering Congress, 2000. He visited Department of Computer Engineering, King Khalid University, Abha, Kingdom of Saudi Arabia as Professor from October 2010 to February 2011 and Department of Computer Engineering and Information Technology, Al-Hussein Bin Talal University, Ma’an, Jordan as Assistant Professor from October 2004 to June 2005. His research interests include image processing, data compression, security protocols, parallel computing, soft computing, and biometrics.

Part I
Cloud Computing, Distributed Systems, Social Networks, and Applications

Review of Elasticsearch Performance Variating the Indexing Methods

Urvi Thacker, Manjusha Pandey and Siddharth S. Rautaray

Abstract In today's world, data is increasing rapidly. Users mostly refer to internet for any information. Also a recent study shows that most of the users go to a search engine to refer to any other site also. So search has become an inseparable activity in internet. Elasticsearch is a java-based search engine that works efficiently in cloud environment. It mainly serves purpose of scalability, real-time search and efficiency that relational databases were not able to address. In this paper, we represent our involvement with Elasticsearch, an open source, Apache Lucene-based, full-text search engine that gives near real-time search ability, as well as a RESTful API for the simplicity of access to users in the various fields like education and research.

Keywords Elasticsearch · Indexing · Restful · nGram · Non-nGram

1 Introduction

This new era is of data. Data is increasing at an unbelievable rate. With the rise in data the difficulty to analyse and search it is also rising. Elasticsearch is a type of database server that is meant for big data search. It can scale to hundreds of servers and petabytes of data (structured/unstructured). It works on the concept of inverted index and is mainly based on Apache Lucene. It can do some other intelligent assignments, but at its core it is made for moving through text, returning text similar to a given query and/or statistical analyses of a collection of text [1]. Elasticsearch additionally permits to consolidate geolocation with full-text search, structured

U. Thacker (✉) · M. Pandey · S.S. Rautaray
School of Computer Engineering, KIIT University, Bhubaneswar, India
e-mail: thackerawake@gmail.com

M. Pandey
e-mail: manjushafcs@kiit.ac.in

S.S. Rautaray
e-mail: siddharthfcs@kiit.ac.in

P.K. Sa et al. (eds.), *Progress in Intelligent Computing Techniques: Theory, Practice, and Applications*, Advances in Intelligent Systems and Computing 719, DOI 10.1007/978-981-10-3376-6_1

search, and analytics [2]. It can be combined to other big data tools like Hadoop, to provide real-time searches for log data [3]. Working with it is very simple and advantageous as its fundamental convention is HTTO/JSON (Java Script Object Notation). It is easy to install. Also the default configuration is sufficient for any new user.

Elasticsearch mainly supports two types of indexing—nGram and non-nGram. NGram is type of indexing in which each field in a document is indexed as strings of length ‘n’. It can be considered as a moving window of length ‘n’ on a word. NGram mainly solves purpose of partial matching. For search type search-as-you-type, a specialized form of nGram called Edge n-grams is used. In non-nGram indexing, fields are stored as it appears in the document.

2 Working Framework

Elasticsearch is a distributed data storage system. It can store and fetch complex data structures serialized as JSON documents in near real time [4]. In other words, the instance in which a document has been indexed in Elasticsearch, it can be retrieved from any node in the cluster. When performing search, JSON objects are given to Elasticsearch and results obtained are also in JSON format. The detailed workflow of Elasticsearch is shown in Fig. 1.

Whenever a user fires a GET or POST query to Elasticsearch, the query gets executed in the system and then the result obtained is returned in JSON format. If there are multiple results then the results are arranged according to their score. Score is a value assigned to each result according to their relevance and is used to arrange the results in presentable format. Elasticsearch uses Lucene’s Practical Scoring Function as default score method. The smallest unit of Elasticsearch is document and largest is index. Shard is an index of Lucene which is the basis of Elasticsearch.

Elasticsearch system automatically detects other system in the network based on the name of the cluster only. If the cluster names of two nodes are same, then they can communicate and need not do any explicit setting for them to connect. In cloud environment network failure is uncertain. So for dealing with network failure Elasticsearch allows nodes to create copy/duplicate of shards called replica shards or replica only. It gives high availability in case any node fails or gets disconnected from the cluster. Replica of a shard and original shard cannot stay on the same node. If original shard is on one then its replicas will be on other nodes.

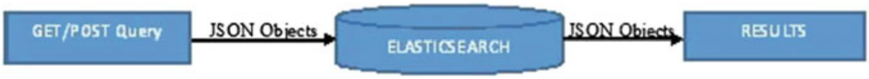


Fig. 1 Workflow of Elasticsearch

3 Dataset

The dataset used for analysis purpose has been taken from internal site of VMware. The dataset has different entities that exist in a virtual environment/help in network virtualization, for example, Logical Switch, Logical Port, Logical Router, Logical Router Port, etc.

4 Implementation

In our experiment, a single node cluster is taken into consideration. Dataset of 1 million documents has been used. At first, indexing of such a huge data is performed and analysed. The CPU usage, heap size, and time for indexing are observed. For indexing purpose, a java programme has been written which run 20 threads for inserting data into Elasticsearch. Then same dataset is analysed for searching. A set of queries with varying result set and page sizes were executed so as to record the time taken to return the results. Also a smaller dataset (200 K) is used for analysing search timings when the data is indexed using nGram and non-nGram.

- *Configuration*: The analysis was carried out on an Ubuntu VM with 8 CPUs and 8 GB memory (IP-10.110.8.5) with Elasticsearch (version 1.7.2 based on Lucene version 4.10.4).
- *Tool used for profiling*: VisualVM for CPU and memory profiling, pidstat for disc I/O.

5 Results and Interpretation

Analysis I
Dataset: 1 million
Search Query Performance: Queries with varying result set (1 million, one-third of 1 million, two-third of 1 million, half of a million and 1) and page sizes (100, 500 and 1000) were executed so as to record the time taken to return the result. As the CPU and memory usage for query execution is very less, we are not taking it into consideration but VisualVM's recordings are given in this document along with the queries.

Index Creation—(Fig. 2)
The graph above shows that the creation of one million entities took almost one and half minutes and CPU usage ranged between 40 and 60%. Heap usage ranged from 100 to 300 MB.

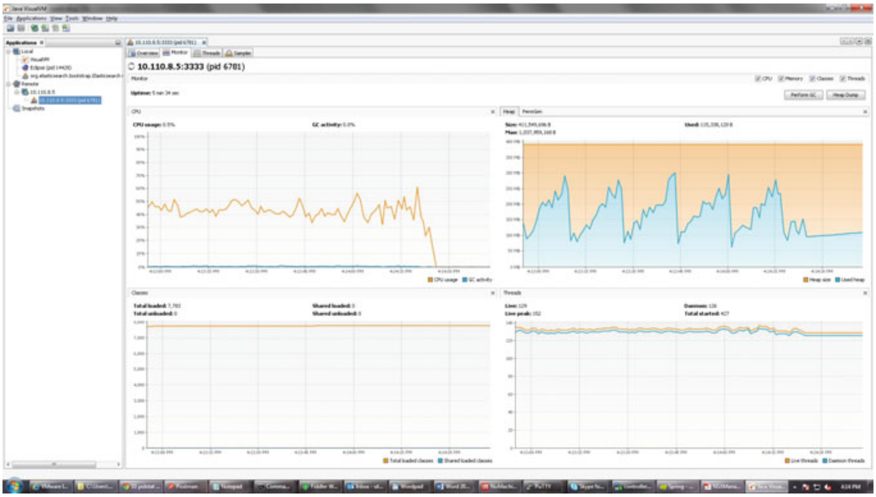


Fig. 2 Index creation of 1 million objects

Table 1 Time of execution of queries with different result counts and page sizes

Sl no	Query	Result count	Time taken (in ms)/page size = 100	Time taken (in ms)/page size = 500	Time taken (in ms)/page size = 1000
1	Query1	1000000	14	25	54
2	Query2	833360	26	85	132
3	Query3	500040	16	27	37
4	Query4	333336	24	30	58
5	Query5	166680	13	24	34
6	Query6	1	3	3	10

Search Queries with different result Count—(Table 1)
The table shows that the result count and page size are directly proportional to the time of execution of queries. But if more complex query is used then for the same result count the time of execution of complex query is higher than the simple one.

Analysis II
Dataset—200 K Objects
Wildcard Query Performance with Index Creation using nGram—A dataset of 200 K entities has been used and time of index creation is taken into account using nGram and non-nGram.
Index Creation—(Fig. 3)
As observed from the figures that when nGram is used CPU utilization is higher and also the time of index creation is higher.

Query Performance—(Table 2)

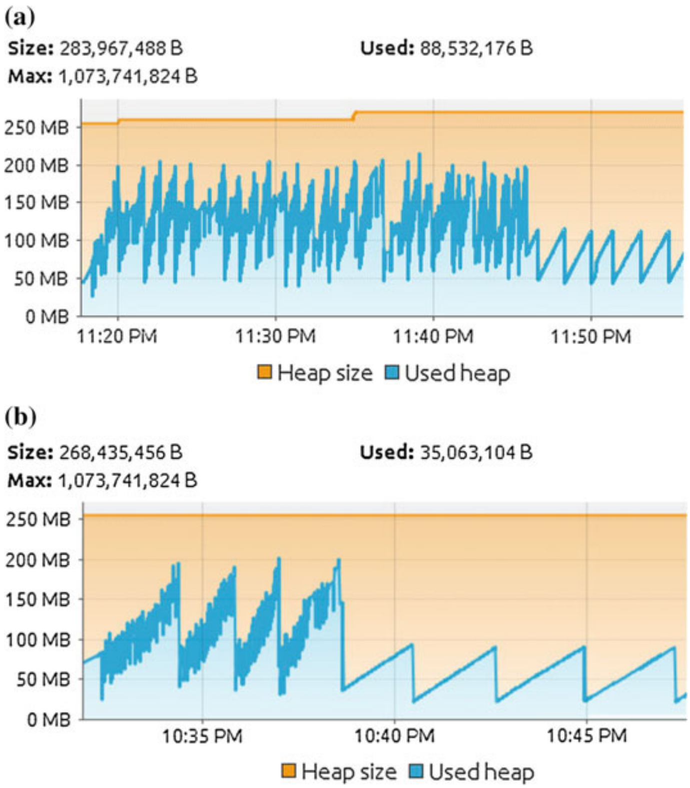


Fig. 3 **a** Index creation of 200 K objects with nGram indexing. **b** Index creation of 200 K objects with non-nGram indexing

Table 2 Query performance in nGram and non-nGram indexing

Use case	Query type	Query	Execution time in ms		
			1st run	2nd run	3rd run

Fetch all documents containing “logical”	Wild card	{
		“query”:{
		“query_string”:{
		“query”:”logical*”
		}
		}
		}
			22	7	5

	NGram index	{
		“query”:{
		“query_string”:{
		“query”:“logical”
		}
		}
		}
			13	7	6

(continued)

Table 2 (continued)

Use case	Query type	Query	Execution time in ms		
			1st run	2nd run	3rd run

Fetch all documents containing “transport OR zone”	Wild card	{
		“query”:{
		“query_string”:{
		“query”:“Transport Zone*”
		}
		}
		}
			19	13	10

	NGram index	{
		“query”:{
		“query_string”:{
		“query”:“Transport (Clinton Gormley, 2015)t Zone”
		}
		}
		}
			194	20	14

References

1. Bai, J. (2013). Feasibility analysis of big log data real time search based on hbase and elasticsearch. *Ninth International Conference on Natural Computation (ICNC)* (pp. 1166–1170). IEEE, Natural Computation (ICNC).
2. Oleksii Kononenko, O. B. (2014). Mining modern repositories with elasticsearch. *Proceedings of the 11th Working Conference on Mining Software Repositories* (pp. 328–331). ACM.
3. Pingkan P. I. Langi, W. W. (2015). An evaluation of Twitter river and Logstash performances as elasticsearch inputs for social media analysis of Twitter. *International Conference on Information & Communication Technology and Systems (ICTS)* (pp. 181–186). Surabaya: IEEE.
4. Tong, C. G. (2015). *Elasticsearch: The Definitive Guide.* O’Reilly Media, Inc.

Understanding Perception of Cache-Based Side-Channel Attack on Cloud Environment

Bharati S. Ainapure, Deven Shah and A. Ananda Rao

Abstract Multitenancy is the biggest advantage of computing, where physical resources are shared among multiple clients. Virtualization facilitates multitenancy with the help of the hypervisor. Cloud providers virtualize the resources like CPU, network interfaces, peripherals, hard drives, and memory using hypervisor. In a virtualization environment, many virtual machines (VMs) can run on the same core with the help of the hypervisor by sharing the resources. The virtual machines (VMs) running on the same core are the target for the malicious or abnormal attacks like side-channel attacks. Cache-based attack in the cloud is one of the side-channel attacks. Cache is one of the resources shared among different VMs on the same core. The attacker can make use cache behavior and can perform the cache-based side-channel attack on the victim. In this paper, we explore different types of cache designs, categories of cache-based side-channel attacks and existing detection and mitigation techniques for cache-based side-channel attacks.

Keywords Hypervisor • Cloud computing • Side-channel attack • CPU cache • Virtual machine • Cache-based side-channel attacks

B.S. Ainapure (✉)
MIT College of Engineering, Pune, Maharashtra, India
e-mail: ainapuressa@gmail.com

D. Shah
Thakur College of Engineering, Mumbai, Maharashtra, India
e-mail: sir.deven@gmail.com

A.A. Rao
JNTU, Anantapur, Andra Pradesh, India
e-mail: akepogu@gmail.com

P.K. Sa et al. (eds.), *Progress in Intelligent Computing Techniques: Theory, Practice, and Applications*, Advances in Intelligent Systems and Computing 719, DOI 10.1007/978-981-10-3376-6_2

1 Introduction

The word cloud computing is sharing of the resources over the internet through different services. Sharing of resources among different users gives multitenancy concept in cloud computing. Multitenancy has both advantages and disadvantages nature. The advantage is that sharing of resources is cost effective for cloud provider and disadvantage is the providing security to users by cloud providers. Multitenancy is implemented on cloud with the help of virtualization. Virtualization is technology which emulates the hardware with the help of software. The virtualization in cloud computing is achieved with the help of virtual machines (VM). Hypervisor is software which creates VMs, manages, and runs in cloud. Hypervisor is also called as virtual machine manager (VMM). With the help of VMMs, more than one virtual machine can be run on a single physical machine that is nothing but host machine. This gives the concept of sharing resources among multiple users, which is nothing but multitenancy. Multitenancy adds new attack surface and new challenges to prevent these attacks. One of the most recent attacks on cloud is side-channel attacks on virtual machines. Side-channel attacks are attacks to steal cryptographic information from hardware resources. Side-channel attacks require knowledge of the internal working of the system hardware. To know the internal working of the system, the attacker has to place his/her VM as a co-resident machine in the cloud environment [1]. The followings are some of the possible side-channel attacks in the cloud environment.

1.1 Fault Attacks

Fault analysis is one of the side-channel attacks. This attack is related to faults in the computation. Through this attack cipher text is examined to extract the keys. Faults are most commonly generated by the change in the voltage or by tampering with the clock.

Fault attacks can categorize in two different categories. The first category of fault attacks is based on computational faults. These types of faults occur during the cryptographic computation. These attacks may be intentional or unintentional. Second category of attacks is intentional. This type of attack is generated by sending corrupted data to the attack module. The following aspects can be considered in fault attack [2]:

- The accuracy of choosing the time and location by an attacker on which the fault occurs during the execution of a cryptographic module.
- Due to fault in computation the length of data may affect.
- The fault may be transient or permanent which shows the property of persistence.
- Errors in flipping the bits in only one direction, which shows the fault type.

1.2 Power Analysis Attack

It is one of the side-channel attacks. This type of attack is performed by analyzing the power consumption of cryptographic hardware device while doing the computations. The power analysis attack is performed on the devices which store secret keys, such as smart cards or any embedded system which store secret key.

Basically, a power analysis attack can be divided into two categories: Simple (SPA) and Differential Power Analysis (DPA). In SPA attacks, the attacker does the guessing of the power traces. Especially in this attack, the attacker does the trace of time and an input/output value of particular instruction is being executed. In such attacks attacker needs to understand implementation details of device to perform the attack. The DPA attack is performed by attacker by exploiting the satirical methods used in the analysis process. A DPA attack does not need the knowledge of the implementation details.

1.3 Electromagnetic (EM) Attacks

The computer is made up of many electrical devices. These devices often generate electromagnetic radiation as part of their operation. Attackers make use of these electromagnetic radiations of device to perform attack. Such type of attacks is called as electromagnetic attacks. In this type of attack, attacker first observes the release of devices and then performs analysis on this release to co-relate their causal relationship to the underlying computation and data. This observation can be inferred to establish relationship between computation and data. Once this information is accessed then the attacker can do anything this with this data.

Similar to the power analysis attacks, electromagnetic analysis (EMA) attacks are categorized as simple electromagnetic analysis (SEMA) and differential electromagnetic analysis (DEMA) [3].

1.4 Acoustic Attacks

Acoustic attack is one of the side-channel attacks, which exploits sounds produced by the computers during the computation. This attack causes the leakage of information about the device and computations occurring on that device. This information leakage can be classified into three different categories [4]: 1. invasive, 2. semi-invasive, and 3. non-invasive.

In invasive attack the attacker tries to get direct access to internal components by tampering the device. In semi-invasive, the attacker gets access to device, but he does not make direct contact instead he makes indirect attack such as fault-based attack. With non-invasive attack, the attacker tries to get externally available information such as sound produced by device while performing some computation, which is unintentionally leaked by devices. The correlation between the sound generated by the processors and its computation is proved for the first in [3].

1.5 Cache-Based Attacks

CPU cache memory is one of the resources which can be shared between multiple virtual machines running on the same host. There are different levels of cache that may be present on processors. If the requested data is not in the cache memory, then cache miss will occur, which causes the main memory reference. Whenever main memory reference occurs, it takes longer time to read the contents. In such situations, the attacker can measure the time taken by the processor to read the cache contents and can do the attack on the cache. This CPU cache is one of the major threats found in cloud computing. This is one of the resources where the cloud provider has to think about security that needs to be provided against cache-based side-channel attacks.

The remainder of this paper is organized as follows. Section 2 talks about cache-based side-channel attacks. In this section we have given an introduction about cache designs, their internal working and different categories of cache-based attacks. We have categorized attacks mainly in two different categories depending on cache behaviors. Section 3 explains about literature review done toward cache-based side-channel attack. Section 4 presents conclusions based on cache attacks.

2 Cache-Based Side-Channel Attacks

2.1 Introduction to the Cache Memory

Cache memory is small memory present in processors. It is also called as CPU cache. It is introduced to match the latency of the memory and speed of the processor. Whenever the processor needs a memory access, the address in the memory that is being reference made is first mapped into cache line. These cache lines are divided into a number of fixed blocks. These blocks are normally 32, 64, or 128 bytes in size. Cache lines are called as rows. Each row has the following structure (Fig. 1).

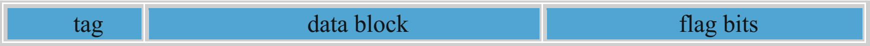


Fig. 1 Structure of cache row

Actual data fetched from the main memory is stored in the data block field of the cache row. Part of the address of the actual data fetched from the main memory is stored in tag field. Flag bits have different usages depending on cache, whether it is instruction or data. Only one bit of flag entry is required per cache row, if cache is instruction cache and that bit is valid or not. Cache block is loaded with valid data, if the bit is set to valid, otherwise at boot time hardware sets all valid bits in all cache lines to invalid. Apart from this processor marks this bit as invalid, when data in the cache block become stale.

Two bits of the flag are required in data cache. These two bits are used per cache line. These are valid bit and dirty bit. Dirty bit on the cache line indicated that the associated cache line has been modified or changes and not yet this change or modified data are written back to the main memory. Otherwise bit is a valid bit.

The cache size normally refers to the actual amount of data in main memory. This size of cache is calculated by the number of bytes stored in each data block times the number of blocks stored in the cache. The effective memory address filed consists of the tag, the index, and the block offset, as shown in Fig. 2.

The index in the effective address defines a cache line that holds the data. The length of the index is calculated as $\log_2$ (L) bits, where L is the number of cache lines. The block offset describes the required data stored within a data block in the cache line. The block offset is calculated as $\log_2$ (B) bits, where B is the number of bytes per data block. The tag length in bits of the effective address is calculated as address_length - index_length - block_offset_length.

For example, to design cache for 32-bit address processors, we may need 16 KB of cache data and each cache block may contain 16 words and the way may adopt 4-way set-associative replacement policy. Then tag size is 20 bits. To get this, the value for tag size considers byte assembler memory, and then each cache block has 16 * 4 bytes equals to 64 bytes which needs 6 bits as the block offset. We know that cache is 4-way set associative and each has 64 bytes of data, and then the index is calculated as 16 KB/(4 * 64) = 2^6 bits. Therefore, 6 bits are needed for indexing. There for tag, length = 32 (address length) − 6 (index length) − 6 (block length).

The mapping of cache blocks with main memory is called as cache associativity. This associativity categories in three different ways: 1. Direct mapped, 2. N-way set associative, and 3. Fully associative.

In direct mapping, each block is mapped to exactly 1 cache line. It is also called as 1-way set-associative cache. In this mapping main memory is divided into number of cache lines. The size of each page in the memory is equal to one cache line. This type of memory mapping is less complex to implement, but it is less flexible to putting the blocks in cache lines. This gives the less performance in switching between the cache lines.

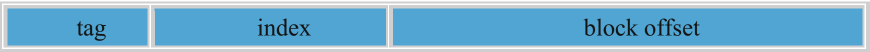


Fig. 2 Effective memory address

In fully associativity cache mapping, each memory block mapped to any cache location. In this type of mapping, size of memory block and cache line are equal. This mapping gives the best performance as any memory block can be stored on any cache line. Implementation wise, it is very complex. It requires a large number of comparators, which increases cost and complexity. Therefore, only cache size less than 4 K is suitable to implement this type of mapping.

Set associativity cache mapping is the combination of fully associative and direct mapping. In this type of mapping, each block is mapped to a subset of cache lines. This mapping is implemented by grouping the cache lines into sets. The number of lines in a set can vary from 2 to 16. The data on this line can be stored in any of the lines in the set. This type of mapping is less complex to implement than fully associative cache mapping as the number of comparators is equal to the number of cache sets. This mapping reduces cost. 2-way, 4-way, or N-way set associativities are also possible in this mapping.

The replacement policy implemented on systems will decide which entry on main memory will go into the cache block. In set-associative and fully associative mapping, the system has to make a decision about where to and what values of the data to be replaced. This decision is made by system by implementing any one of following replacement algorithms.

1. First-in first-out replacement algorithm
2. Least recently used algorithm or
3. Random replacement algorithm.

The block that has been in the cache for a long time is removed from cache in first-in first-out replacement algorithm. This type of algorithm is suitable for set-associative and fully associative mapping. More easily it can be implemented using counters on hardware as it requires only one counter per cache line.

In the least recently used (LRU) algorithm, the cache block which is not referred to the longer time is removed from the cache. This algorithm is most suitable for implementing the 2-way set-associative mapping as it is suitable for small numbers of cache blocks. This type of algorithm is implemented in hardware with the help of counters and register stacks.

In random replacement algorithm, the cache block to be replaced selected randomly without any concern of memory reference or previous selection. Counter is used to implement this algorithm on hardware.

The power consumption and timing of the device mainly depend upon the memory access, which shows the number of successful hits to the memory. During memory block access cache miss can be classified as followings.

Cold start misses: This type of cache miss occurs, when the first access to a block is not in the cache. Then the block must be brought into the cache. So it is called as cold start misses.

Capacity misses: This type of cache miss occurs when cache lines cannot contain all the blocks needed during execution of a program. Capacity misses occur due to blocks being discarded and retrieved later on when required.

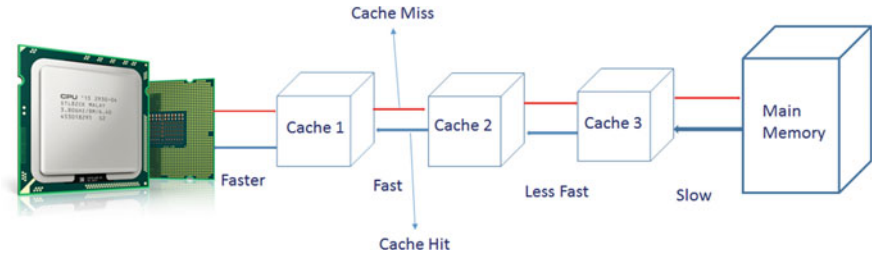


Fig. 3 Hierarchical cache memory

Conflict misses: This type of cache miss is called as collision miss. Normally this type of cache miss occurs in associative or direct mapped cache configuration. The conflict in cache miss occurs, when blocks are discarded and retrieve later, if too many blocks map to its set.

Cache memory may be shared by different processors or may have individual caches. Latest processors have multiple levels of cache. For example, the Core i7 processor has three levels of cache memory for different purposes. When there are multiple levels of the cache normally L1 level of cache is divided into data lines and instruction lines. The hierarchical structure of cache memory is shown in Fig. 3.

Whenever data or instructions are referred by the CPU, it first requests to L1 level cache. If the requested data is found in L1 by CPU, then it is called as a cache hit. If it is not found in L1 cache, then it is called as cache miss. If a cache miss is experienced, then the data are next looked for in the next level of memory—for an L1 cache miss, this would be the L2 cache. If found data is propagated back through each level of cache that experienced, miss to locate the data on the cache and this process of cache hit and cache miss is continued with the next level of caches. Normally, L2 cache is much larger in size than L1 and L3 is larger than L2. L3 is the last level of cache, which stores data from multiple cores simultaneously. If a cache miss at the L3 cache occurs, the requested data are sought in main memory.

To use the CPU cache as side channel to do the attack, some or the other way cache needs to share between attacker and target. On the cloud environment, it is very easy to share the cache between the different VMs launched on the same CPU core. Typically, a CPU cache can be shared in two ways, either the cache is exclusive to one CPU core, in which case two processes must access the cache sequentially or cache is shared between CPU cores, in which case two processes can access the cache concurrently.

2.2 Types of Cache-Based Side-Channel Attack

Cache attacks can be categorized in different ways. Broadly they can be categorized based on microarchitecture attacks and based on cache miss during cryptography.

2.2.1 Based on Microarchitecture Attack

According to microarchitecture attack, attacks can be categorized in three different ways:

1. Access-driven attack
2. Time-driven attack
3. Trace-driven attack

Access-Driven Attacks

In access-driven attacks, the attacker gets the information about the victim's cache by accessing the common shared cache. The cache can be accessed in sequential manner or in parallel based on whether concurrent access is made or sequential access is made on the cache [5]. During this type of attack, the attacker can observe instruction cache [6], data cache [1], and branch prediction cache [7] to get the information about the cache sets. This type of access-driven attacks is performed with the help of a technique called as prime and probe. During prime stage attacker fills the cache lines with his/her own data. Then the attacker will wait for predefined time for the victim to access the cache lines. After waiting for predefined time the attacker starts with probe stage, which refills the same cache lines with attacker's data and simultaneously monitors the victim's activity on cache lines. During this stage if victim accesses primed line, then cache miss will cause. This causes more time to read the cache line for victims, which is observed by the attacker to calculate the time span and guess the data to read from the main memory.

Flush and reload cache attack is another variant of the prime probe attack. This type of attack is performed on the last level of cache [7]. In this type of attack the spy program is planted in the environment to do the attack. This attack happens in three stages. During the first stage the monitored memory lines are flushed from cache hierarchy, then spy is allowed to wait for the victim to access the memory line before third stage. During the third stage the spy reloads the memory line and it will measure the time needed to load the cache. During the wait time if the victim accesses the memory line, the line is available with cache, then reload will take less time and if it is not available, then reload will take quite long time to load it [7]. This time is noted by the spy to guess the memory operation carried out by the victim.

Time-Driven Attacks

In time-driven attacks, the attacker tries to measure the total execution time of the cryptographic operations. The total execution time is dependent on a key value, which accesses the memory results in cache misses. In time-driven attack, lots of statistical analyses are required to infer the key value. Evict and prime or prime and probe are the type of time-driven attacks. Attack on AES encryption algorithm can be done with the help evict and prime [8] by evicting in the cache and measuring the time for encryption. In such attacks, attacker do various rounds of encryption and evicts one of the selected cache lines by writing his/her own data and measures the time taken by the victim to do encryption. Then for a victim encryption time depends on the values present in the cache at the starting time. If the victim accesses the evicted cache, then obviously victim will go to take longer time to do the encryption, as cache miss occurs [9].

Trace-Driven Attacks

Trace-driven attacks need knowledge about the underlying hardware and its implementation [10]. In this attack, the attacker keeps track of victim's cache activities. The attacker keeps track of total number of cache lines used and cache miss ratio during encryption round for victim machine [11]. To make these attacks more powerful the attackers have to continuously monitor the CPUs caches. Sometimes trace-driven attacks are based on power analysis, physical access of memory, and alteration of the processing device [11].

2.2.2 Based on Usage of the Cache Miss During Cryptographic Algorithms

During cryptographic operations like encryption or decryption, behavior of cache memory depends on its initial state and subsequent sequence of memory accesses. The attacker can make use of this cache behavior to target the cryptographic devices. Therefore, based on the initial state of cache, attacks can be categorized as follows [12].

Empty initial state: This is also called as reset attack. This type of attack is performed based on the observation of cold start misses. This type of attack does not require lookup tables stored in cache by encryption algorithms.

Forged initial state: These types of attacks are also called as initialization attacks. In these types of attacks, attacker generates chosen number of cold misses to the known state of cache before encryption.

Loaded initial state: In this state of attack the attacker needs cache which is already filled by all the lookup tables involved in the encryption algorithm. In this attack, for a given initial state the sequence of memory accesses performed during the encryption can be observed by timing analysis or by power analysis as suggested in [12].

3 Related Work

This section will discuss about the existing security techniques presented by the researchers most recently toward securing the cloud environment. The literature which will talk about the cache-based side channel is categorized in three different ways that is based on attacking techniques, detection techniques, and mitigation techniques as follows:

3.1 Existing Cache-Based Attacking Techniques

Cache-based side-channel attack is not new concept in literature. This type of attack is possible in the cloud due to resource sharing. For the first time in 2009, cache-based side-channel attacks were proved on the Amazon EC2 cloud environment. This is proved by placing the attacker machine as co-resident with victim's machine on cloud [1]. In this paper author used prime, probe, and trigger technique to extract the cache contents of the victim's machine. But in this paper authors have not proposed mitigation solution. Symmetric multi-processors are used to prove the cache-based side-channel attack using an ElGamal decryption key with the libgcrypt cryptographic library. The authors have used support vector machine to classify the attacks and also applied hidden Markova chain model to reduce the errors in the cache. But they have specified about any specific mitigation technique to handle such types of cache-based attacks [13].

Memory bus contention was used to do a cache-based attack on the victims. This attack was proved on × 86 machine by exploiting the instruction level cache. This attack was performed with the help of some hardware modification which was resulted in overhead in processors and even mitigation techniques were also not proposed [14].

3.2 Existing Techniques for Detection of Cache-Based Attacks

Two-stage detection method is used to detect cache-based side-channel attack. Two stages consist of host and fast detection. The shape test and regularity tests are used to extract attack features from the host and guest, and they also used a pattern recognition technique to differentiate between legitimate VM and attacker VM [15].

The machine-learning technique is applied to detect cache-based side-channel attack. Authors have specially focused on flush and reload method to detect the spy planted on victim through cache attack. To detect the attacks neural network is used with supervised mode [16].

3.3 Existing Techniques to Mitigate Cache-Based Side-Channel Attacks

Some of mitigation techniques were proposed on L2 cache-based side-channel attacks. Different bystander workloads are used for cross-VM covert channels to detect the cache attack and on this continuous time, Markov Process to model was used to mitigate the classified attacks [17]. Threshold values are set between two VCPUs for overlapped scheduling to mitigate the L2 cache attack. In this paper authors have also introduced some noise function to reduce the attack so that attacker cannot get the exact information [18].

Another mitigation technique is proposed based on instruction execution at each user level thread on a single CPU core. In this work authors are not using cache flushing rather they maintain integrity of the data [19].

L3 cache attacks are handled by managing locks on cache lines. This work little bit modification is done on the existing hardware to lock the cache lines. Never evicted cache lines are used to lock so that they can be multiplexed for VM to load their data [20]. But overhead is incurred due to system-level changes.

The cache coloring technique is used to partition the cache on the VM level to prevent the cache-based side-channel attacks. Apart from this firewall security is applied to filter the unwanted requests. But authors have not mentioned authentication levels for users. Overhead is introduced by applying both security aspects [21].

A dynamic page coloring method called as Chameleon is used to prevent cache-based side-channel attack. Chameleon provides stringent isolation between security critical operations and normal operations by assigning a specific color to the process through dynamic page coloring. The dynamic page coloring notifies the hypervisor for entering into the critical section by providing a specific interface for applications [22].

4 Conclusion

Cloud computing has evolved in only in the last decade and still there are numerous of users who are hesitant to adopt the service owing to the unreliability factor of security on the cloud. One of the security threats found in the cloud is side-channel attacks due to multitenancy. The cache-based side channel is not a new concept, but it has gained focus in a cloud environment as resources like cache are shared among multiple users. In this paper, we surveyed papers related to cache-related side-channel attacks on a cloud. Few papers only talk about attaching methods adopted on cloud and few papers talk only about a particular mitigation method. Most of the papers adopted the cache flushing as mitigation. But this cache flushing is not always cost effective as it will pay some penalty.

In this prospect, there is a need for a new framework, which can identify the pattern of all cache-based side-channel attacks and can give the mitigation techniques without any overhead and penalty.

References

1. T. Ristenpart, E. Tromer, H. Shacham, and S. Savage, "Hey, you, get off of my cloud: exploring information leakage in third-party compute clouds," in *CCS*, 2009, pp. 199–212.
2. Side-Channel Attacks: Ten Years after Its Publication and the Impacts on Cryptographic Module Security Testing YongBin Zhou, DengGuo Feng State Key Laboratory of Information Security, Institute of Software, Chinese Academy of Sciences, Beijing, 100080, China.
3. A. Shamir, E. Tramer. Acoustic cryptanalysis: on nosy people and noisy machines. Eurocrypt 2004 rump session, 2004.
4. Deepa G M et al, "An overview of acoustic side channel attack", International Journal of Computer Science & Communication Networks, Vol 3(1), 15–20.
5. Godfrey, M., Zulkernine, M., "Preventing cache based side channel attacks in a cloud environment", IEEE Transactions on Cloud Computing, Volume: 2, Issue: 4, Oct.-Dec. 1 2014.
6. Acıiçmez, O., Brumley, B. & Grabher, P., 2010. New results on instruction cache attacks. In CHES'10 Proceedings of the 12th international conference on Cryptographic hardware and embedded systems.
7. Yuval Yarom, Katrina Falkner "FLUSH + RELOAD: a High Resolution Low noise, L3 cache Side-Channel attack", 23rd USENIX Security Symposium (USENIX Security 14) (San Diego, CA, Aug. 2014), USENIX Association, pp. 719–732.
8. Osvik, D., Shamir, A. & Tromer, E., 2006. Cache attacks and countermeasures: the case of AES. In Topics in Cryptology–CT-RSA 2006. pp. 1–25.
9. Liu, F. & Lee, R.B., 2013. Security testing of a secure cache design. In Proceedings of the 2nd International Workshop on Hardware and Architectural Support for Security and Privacy - HASP'13. New York, New.
10. Raphael Spreitzer, Benoıt Gerard, "Towards More Practical Time-Driven Cache Attacks" WISTP 2014, LNCS 8501, pp. 24–39, 2014.
11. K. Tiri, O. Acıic¸mez, M. Neve, and F. Andersen, "An analytical model for time-driven cache attacks," in FSE'07, ser. LNCS, A. Biryukov, Ed., vol. 4593. Springer, 2007, pp. 399–413.
12. Anne Canteaut, Cedric Lauradoux, and Andre Seznec. Understanding cache attacks. Technical Report, April 2006. Available at: ftp://inria.fr/INRIA/publication/publi-pdf/RR/RR-5881.pdf.
13. Fei Liu, Lanfang Ren, Hongatao Bai, "Mitigating Cross-VM Side Channel Attack on Multiple Tenants Cloud Platform", i-scholar, Journals of computers, Vol. 9 2014, Published: 2014-04-01, pp: 1005–1013.
14. Z. Wu, Z. Xu, and H. Wang, "Whispers in the hyper-space: High-speed covert channel attacks in the cloud," in *USENIX Security*, 2012, pp. 9–9.
15. Si Yu, Xiaolin Gui, Jiancai Lin, "An approach with two stage mode to detect cache based side channel attack", IEEE Computer Society Washington, DC, USA ©2013.
16. M. Chiappetta, E. Savas, and C. Yilmaz, "Real time detection of cachebased side-channel attacks using hardware performance counters," Cryptology ePrint Archive, Report 2015/1034, 2015, http://eprint.iacr.org/.
17. Rui Zhang, Xiaojun Su & et al, "On mitigating the Risk of Cross-VM Covert Channel in Public Cloud:", Parallel and Distributed Systems, IEEE Transactions on (Volume:26, Issue: 8), Date of Current Version: 13 July 2015 Page(s): 2327–2339.

18. Fei Liu, Lanfang Ren, Hongatao Bai, "Mitigating Cross-VM Side Channel Attack on Multiple Tenants Cloud Platform", i-scholar, Journals of computers, Vol. 9 2014, Published: 2014-04-01Pages: 1005–1013.
19. Deian Stefan, Pablo Buiras, & et al. "Eliminating Cache-Based Timing Attacks with Instruction-Based Scheduling", 18th European Symposium on Research in Computer Security, Egham, UK, September 9–13, 2013. Proceedings, Publisher Springer Berlin Heidelberg.
20. Taesoo Kim, Marcus Peinado, Gloria Mainar-Ruiz, "STEALTHMEM: System-Level Protection Against Cache-Based Side Channel Attacks in the Cloud", Security'12 Proceedings of the 21st USENIX conference on Security symposium" ACM, 2012.
21. Godfrey Zulkernine M, "Preventing Cache-Based Side-Channel Attacks in a Cloud Environment", Cloud Computing, IEEE Transactions on (Volume:2, Issue: 4), Issue Date: Oct.-Dec. 1 2014, Page(s): 395–408.
22. Jicheng Shi, Xiang Song, Haibo Chen, Binyu Zang "Limiting cache based side channel in multi-tenant cloud using dynamic page colouring", 2011 IEEE/IFIP 41st International Conference on Dependable Systems and Networks Workshops (DSN-W).

A Semantic Approach to Classifying Twitter Users

Rohit John Joseph, Prateek Narendra, Jashan Shetty and Nagamma Patil

Abstract Social media has grown rapidly in the past several years. Twitter in particular has seen a significant rise in its user audience because of the short and compact Tweet concept (140 characters). As more users come on board, it provides a large market for companies to advertise and find prospective customers by classifying users into different market categories. Traditional classification methods use TF–IDF and bag of words concept as the feature vector which inevitably is of large dimensions. In this paper we propose a method to improve the method of classification using semantic information to reduce dimensions of the feature vectors and validate this method by feeding them into multiple learning algorithms and evaluating the results.

Keywords Semantic analysis · Social media analysis · Named entity recognition

R.J. Joseph (✉) · P. Narendra · J. Shetty · N. Patil
Information Technology Department, National Institute of Technology
Karnataka Surathkal, Mangalore 575025, India
e-mail: rohitjjoseph@gmail.com
URL: http://infotech.nitk.ac.in/

P. Narendra
e-mail: prateeknar@gmail.com

J. Shetty
e-mail: jashanshetty94@gmail.com

N. Patil
e-mail: nagammapatil@nitk.ac.in

P.K. Sa et al. (eds.), *Progress in Intelligent Computing Techniques: Theory, Practice, and Applications*, Advances in Intelligent Systems and Computing 719, DOI 10.1007/978-981-10-3376-6_3

1 Introduction

Social media usage has seen a meteoric rise in the past few years with people using it on a daily basis to share information varying from important announcements to the meal they just finished. Twitter in particular has been one of the driving forces, primarily because of the simplicity of its concept. The 140 character limit is reminiscent of a messaging service, something that can be used to share one's thoughts without batting an eyelid. It is for this reason that around 350,000 Tweets are sent every minute. Twitter is often used by users to provide graphical, textual or a web-link content and due to the unrestricted nature of the content posted by each user, it provides an ideal platform to learn about a person. This information learnt can be used to further enhance the user's experience on the platform by recommending people of similar interests or it can be used when there is a need to target-specific audiences.

While recommendation systems are common places in social networks, most of these recommendation systems are based on the graph structure of existing social networks. A user is categorized into various categories based on the sets of user followed. This is extensive work done on this model, with similarity of users computed using cosine similarity functions and Euclidean distance mechanisms. However, identifying similar sets of users based on the content tweeted is far more challenging. There exists various sources for information in the tweet, the URL or link can be analysed to predict the nature of the article shared and thus the category to which the tweet belongs. The nature of tweets can also be studied by analysing the text present in the tweet. Sophisticated computer vision algorithms can also be used to categorize the image shared by the user.

2 Literature Survey

Following Barbier et al. [1], we see the different data mining techniques required to analyse and gather information from large social media data which is continuously changing. With the help of Derczynski et al. [2] we address the problem of part-of-speech tagging. These aforementioned methods are of prime importance to us, since semantic analysis of the data requires us to first distinguish among the various entities in the text, i.e., nouns, as they convey the most information.

Hannon et al. [3] propose a technique which utilizes the content tweeted by the user, as mentioned above, along with the list of people he/she follows in order to recommend users to one another. While this is an interesting and unique approach, the authors use the TF–IDF weighting metric in order to build a profile for a user, which does not take into context the meaning or sentiment in the text.

Semantically analysing tweets involves identifying the context of every word in the tweet and then trying to use this information in order to classify the user according to the content tweeted. This is a far more complex method to analyse a user than based on the graph structure of the user. A method of semantic analysis of Twitter data has been proposed by Agarwal et al. [4].

A semantic approach we propose involves building a feature list for every user, building a taxonomy for every word in a users feature list, analysing the context of every word in the feature list, assigning a score to ‘k’ parameters, where ‘k’ is the generic categories under which all words can be accommodated and finally feeding this score vector to learning algorithms to group similar users.

3 Proposed Approach

Twitter has taken down many of the publicly available data sets, and as a result of this, data collection (around 30k tweets) had to be done manually using the Twitter Python API, Tweepy [5]. This data is in a raw, unstructured form and needs to be thoroughly cleaned before it can be used to extract useful entities in the preprocessing phase. This is a pretty standard operation in any data mining venture and for this we use the NLTK [6] library along with a dictionary of commonly used Twitter abbreviations to remove unnecessary spam words as well punctuations and so on.

Hashtags are commonly used in Twitter to signify that a particular topic is being spoken about. For example, a user talking about a particular sporting event may end his Tweet with the hashtag, ‘#sport’. This information must also be taken into consideration as it can provide information of interest to the user. Similarly, during the preprocessing phase emoticons, a metacommunicative representation of a facial expression must also be extracted because they provide information in ascertaining the emotion behind a particular statement. This will further help the classification process.

With this initial stage of preprocessing done, we have removed unnecessary symbols and stopwords. The remaining words need to be categorized into various parts of speech (POS), and nouns will be useful for determining the topics of interest to a person while verbs need to be removed. Because different words can take different forms in the English language, it is important to segregate the nouns from the verbs in this process. For our purpose, we evaluated two POS taggers, Stanford POS Tagger and CMU POS Tagger, both of which are written in Java. On closer analysis we found that the CMU POS Tagger was significantly faster than the Stanford POS Tagger. The CMU Tagger offered 220 tweets/sec as opposed to the Stanford 3 tweets/sec while both offered similar accuracies. The disparity may be because the CMU tagger was trained on Twitter data and so performed better.

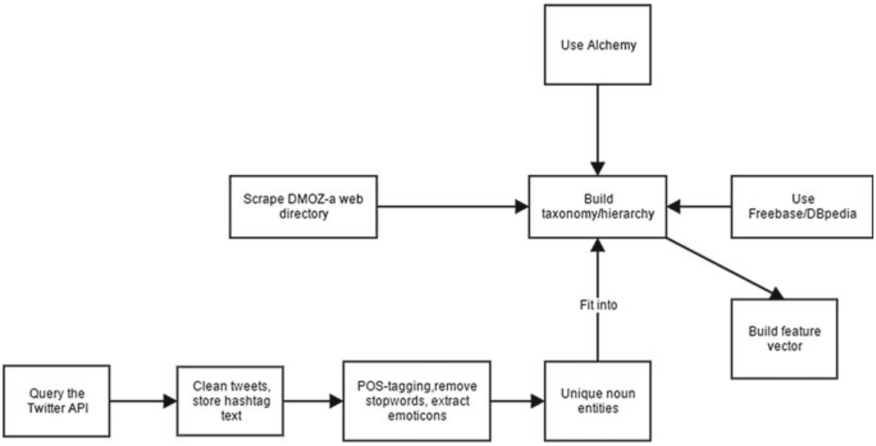


Fig. 1 Workflow

After the thorough preprocessing stage, we are left with the noun phrases as well as the emoticons associated with the corresponding text. Traditional approaches use these bags of words along with a TF–IDF score corresponding to each word as the feature vector associated with a particular user. However, as the number of unique words per user increases, the feature vector size also becomes too large and cannot be fed directly into any learning algorithm, and some sort of dimensionality reduction needs to be done. The complete workflow of proposed approach is comprehensively illustrated in Fig. 1.

3.1 Building the Feature Vector

To solve the problem of the increasing sizes of the feature vector formed by representing the user by the unique words in his text, we propose an approach to build the user feature vector by understanding the semantics of the words in his/her profile. Take for example the football team Manchester United; we know that it is a soccer team and we know that soccer is a sport, so there should be some way in which we can link Manchester United back to its broad category, i.e. sport. Since we have an idea of the broad categories of the data within our dataset, we can build a taxonomy with these categories as the 'root' and various sub-categories under it, as shown in Fig. 2. The taxonomy was built by scraping DMOZ [7], which is the largest directory of human edited information on the web.

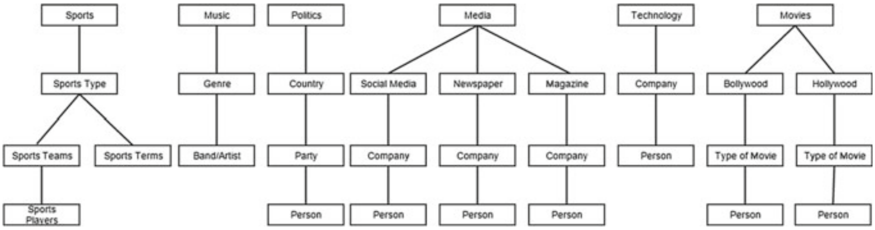


Fig. 2 Taxonomy

For every word that has been tweeted so far, we now try to fit it into this hierarchy. For this we make use of the Alchemy API [8]. Alchemy is a web-based REST-API which provides us with entities, sentiment score and entity-wise sentiment for a given article. The entities obtained in this step are mapped onto an existing database of entities we have created by crawling some of the open web dictionaries [9] available for the taxonomy that we had created. The entities for which a match is not found in the list are discarded. A score for each of the broad categories, as shown for five users in Table 1, is then calculated by taking the weighted average of the number of matches of a word within a particular category's subtree along with a bias weight given according to the sentiment associated with the text (also taking into consideration the emoticons extracted earlier).

4 Results and Discussion

As shown in Table 1, the feature vector corresponding to a particular user has been reduced from around 150 on an average (average number of unique words tweeted per user) to 6. The scores assigned have been calculated by mapping each word in the tweet onto the level they correspond on the hierarchy. All the five users shown in the figure are journalists or people who report about different sporting topics, which is seen with their highest score coming in the 'sports' category. For this paper, we have tried to classify the users into the six broad categories at the top of the hierarchy.

Table 1 Reduced feature vector

Name	Sports	Music	Movies	Politics	Technology	Media
Henrywinter	0.12	0.057	0	0.057	0	0.017
Bumblecricket	0.081	0.01	0	0	0	0.009
Samwallacetel	0.19	0.01	0	0.005	0	0.019
Marcotti	0.068	0	0	0	0	0.023

Table 2 Evaluating the proposed method

Learning algorithm	Precision	Recall
Support vector machine	0.85	0.84
Logistic regression	0.74	0.79

We fed this into two classification algorithms, namely SVM and multiclass logistic regression (Table 2). Since we have a nonlinear data set, we used nonlinear SVM with different kernels. The metrics we have used are
a) Precision
Precision is the True Positive value divided by the sum of True Positive and False Positive

$$Precision = \frac{tp}{tp + fp} \tag{1}$$

b) Recall
Recall is the True Positive value divided by the sum of True Positive and False Negative

$$Recall = \frac{tp}{tp + fn} \tag{2}$$

5 Conclusions and Future Work

Semantically, building a feature vector helps reduce the dimensions of the vector without losing information in the process. Rather, it can be exploited to reflect a lot of information of the underlying text. By considering the sentiment along with the semantic meaning of the text present in the Tweet, it greatly helps reduce the dimensions of the feature vector, thus enabling the machine learning algorithm to run quicker while still maintaining a high level of accuracy.

We have limited our feature vector to be of six dimensions and thereby built our taxonomy in that way. In the future we look to expand this, to increase the number of categories which we map words on to and in this way further increase the number of classes we classify our users into. Another scope of improvement would be to improve the method of score calculation. A nonlinear function of the occurrence count may further improve results. Also, we look to exploit the underlying graph structure of a social network and use collaborative filtering techniques mentioned by Hannon et al. in [3] to further improve our results.

References

1. Barbier, G. and Liu, H.: *Data mining in social media*. In: Social Network Data Analytics, pp. 327–352 (2011)
2. Derczynski, L., Ritter, A., Clark, S., Bontcheva, K.: Twitter Part-of-Speech Tagging for All: Overcoming Sparse and Noisy Data. In: Proceedings of Recent Advances in Natural Language Processing (RANLP). Association for Computational Linguistics (2013)
3. Hannon, J., Bennett, M., Smyth, B.: Recommending twitter users to follow using content and collaborative filtering approaches. In: Proceedings of the fourth ACM conference on Recommender systems, pp. 199–206. ACM (2010)
4. Agarwal, A., Xie, B., Vovsha, I., Rambow, O., Passonneau, R.,: Sentiment analysis of Twitter data. In: Proc. ACL 2011 Workshop on Languages in Social Media, pp. 30–38 (2011)
5. Tweepy.org, http://tweepy.readthedocs.org/en/v3.2.0/
6. Natural Language Toolkit, http://www.nltk.org
7. DMOZ, https://www.dmoz.org/
8. AlchemyApi, http://www.alchemyapi.com/
9. DBpedia, http://wiki.dbpedia.org/

Timeline-Based Cloud Event Reconstruction Framework for Virtual Machine Artifacts

B.K.S.P. Kumar Raju and G. Geethakumari

Abstract Traditionally, scaling the resources to meet the high dynamic needs of consumers is a challenge for organizations. Alongside cost, maintenance overheads, availability issues are contemplated. A scientific solution that considers all these is *cloud computing*. Moreover, recent advancements in cloud also allured many small and medium scale enterprises. But, the extent of security and privacy provided to the tenant's data is not apparent and proper. Contemporary attacks on the cloud strengthen this argument. A reactive approach to handle the occurred incident in cloud is through performing forensics. But the domain of *cloud forensics* is still in its infancy state. In mid 2014, National Institute of Standards and Technology (NIST) released a draft which contains various legal, organizational, architecture, and *technical challenges* to perform forensics in the cloud environment. In this paper, our focus is on one of the technical challenges namely *Event Reconstruction*. We consider cloud virtual machine artifacts to achieve the same.

Keywords Cloud computing · Digital forensics · Reconstruction · Event correlation

1 Introduction

Cloud computing is an on-demand computing environment where the resources are shared by ubiquitous access. Its advantages have had high appeal for both the industry and academia in less time. But the lack of transparency posed by the cloud is becoming a setback in using it for sensitive and secured applications. A lot of recent attacks strengthen this argument. Few examples are, DDoS attack has been performed on amazon.com in 2011 [1]. In 2013, hackers have used Dropbox to perform advanced persistent threats [2]. In 2015, from amazon EC2 co-located instance the

B.K.S.P. Kumar Raju (✉) · G. Geethakumari
BITS Pilani Hyderabad Campus, Hyderabad 500078, India
e-mail: pavan0712@gmail.com

G. Geethakumari
e-mail: geetha@hyderabad.bits-pilani.ac.in

P.K. Sa et al. (eds.), *Progress in Intelligent Computing Techniques: Theory, Practice, and Applications*, Advances in Intelligent Systems
and Computing 719, DOI 10.1007/978-981-10-3376-6_4

RSA decryption keys were recovered which violated the privacy of other users [3]. Recently, students at Worcester Polytechnic Institute hacked an instance in AWS which lead to a data breach. Similar attacks may still continue in the future if the underlying isolation and cache management policies are not improved at both the hardware and software level.

To address this, the provider has to enhance the trust of users on the cloud environment which can be done in two possible ways: 1. Developing a highly secured policies and algorithms. 2. Performing forensic investigation in the cloud with higher transparency. In this paper, we focus on the latter. There are no standard techniques for performing cloud forensic investigation as this is a relatively new area. Moreover, traditional digital forensic techniques cannot be directly applied to the cloud environment because of its architecture and unique characteristics. Giving the same conclusion, NIST working group had identified 65 challenges in performing forensics for the cloud and segregated them into organizational, technical, architecture and legal issues [4].

In this paper, we emphasize on one of the technical challenges namely *event reconstruction*. There are many definitions for event reconstruction but the most standard one is given by [5] i.e. *Process allowing to describe an incident using information left on a crime scene*. Event reconstruction would give many advantages: (a) It helps the investigator to generate the hypothesis (b) It is used to examine the evidence and know why it possess certain characteristics (c) It increases the interpretation of the overall crime scene under consideration.

In Sect. 2, we present the related work for event reconstruction. A framework for cloud-based event reconstruction was proposed in Sect. 3. The proposed framework is then validated with the corresponding experiments in Sect. 4. Finally we summarize the work in Sect. 5.

2 Related Work

We present the work in event reconstruction for both the cloud and non-cloud environments. Irrespective of the environment, a thumb rule is that event reconstruction can only answer simple questions but may not answer complex ones in all the cases [6].

2.1 Cloud-Based Event Reconstruction

Deleted data act as a crucial evidence as it may give many conclusions about the incident [7]. Recovering the deleted data in the cloud environment is much more challenging due to its multi-tenant nature. With the aid of quick snapshot feature of cloud, there are chances to get the deleted data. But, performing event reconstruction using the deleted data is challenging.

In [8], the authors claimed that by maintaining proper backups, the process of event reconstruction may be possible. But in reality, this has to consider: the cost involved in having the backups, frequency of backups, extent of data to backup, etc. Moreover, due to the distributed nature of cloud, there is a need to do time synchronization. An improper synchronization would not lead to a fool proof reconstruction of events.

Recent papers commented on the possibilities and challenges of performing cloud-based event reconstruction:

- Reconstructing the virtual storage from the physical cloud node is still a challenge [9].
- Event reconstruction in cloud is tricky as the data needs to be considered from multiple sources. Moreover, finding the relevant data pertaining to the incident is also challenging [6].
- There may be multiple versions of the same file at same time which may increase the complications in event reconstruction [10].
- Cloud-based event reconstruction may be possible if the subset of the environment is considered but not the whole environment [6].

2.2 Non-cloud-Based Event Reconstruction

We can categorize the work in traditional event reconstruction as timeline based and not completely timeline based (i.e. non-timeline based). In this paper, our emphasis is on timeline-based event reconstruction and the work related to it is presented below:

Events containing time as the parameter are either extracted from the filesystem or from the inside files. There are few tools like sleutkit [11] and Encase [12] that generate timeline from the target filesystem. For extracting the time events from each file, various prototypes and tools exist like Cyber Forensic Time Lab (CFTL) [13], log2timeline [14], Zeitline [15]. Most of these tools suffer with proper visualization of the events especially when the number of events is more. This becomes worse, when we analyze the cloud artifacts as they may have more number of events than the traditional environment artifacts.

So our objective in this paper is to reduce the number of events for timeline generation without much data loss of highly important events. To the best of our knowledge, the existing timeline approaches have considered events from only one artifact at a time. But a combined timeline of all the artifacts would give more comprehensive view about the system state which we are considering and addressing in this paper. We examine three major cloud virtual machine (VM) artifacts for timeline generation-vRAM, vDisk and cloud service logs. We identified various challenges when multiple artifacts are considered for timeline like data loss, segregation, numerous events, time synchronization, unique format, etc.

3 Framework for Cloud-Based Event Reconstruction (CERf)

In general, *Event reconstruction* assists the investigator in many ways to reflect various advantages like: increases the admissibility of the evidence in the court of law, reduces the possibility of erroneous logical conclusions, and even it reduces the time for the investigation by enhancing the interpretation of the underlying incident.

Unfortunately, the research on cloud-based event reconstruction is at its early stages and is facing several challenges as mentioned in Sect. 2. So, we proposed a framework for performing event reconstruction in the cloud environment (Fig. 1). The base for the proposed framework is taken from the work done in *traditional digital event reconstruction* [16].

3.1 Phases for Cloud Event Reconstruction

Our framework contains *new phases* to suit for cloud environment and they are briefed as below:

- Segregation: Since the cloud is a multi-tenant environment, the investigation on the target VM artifacts (vRAM, Service logs, vDisk etc.) should not violate the privacy of other tenants. For example, consider the artifact namely *service logs*. These are present within the cloud nodes and they contain events of all the cloud

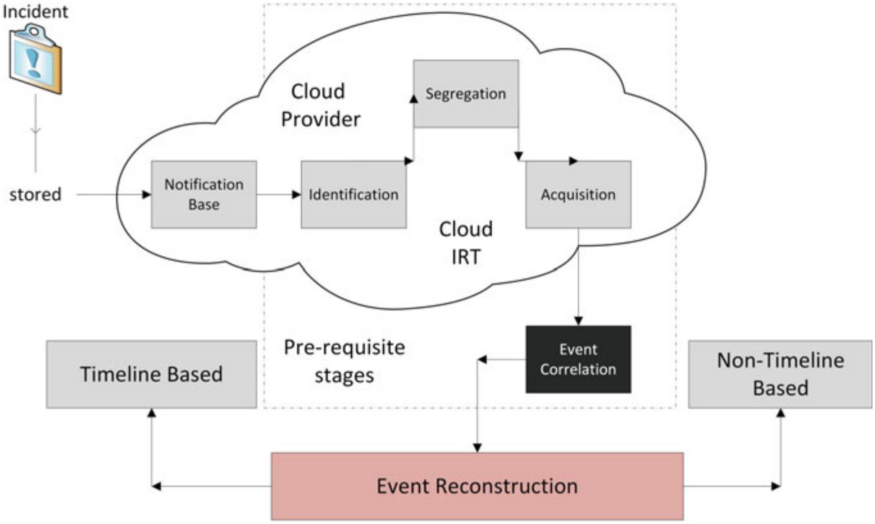


Fig. 1 Proposed framework for cloud event reconstruction

users. If CSP gives all these logs to the investigator who may be interested in only a specific VM's events then it violates the privacy of other tenants (users). So, segregation acts as a prerequisite step before the actual acquisition phase, i.e., this phase defines the segregation parameters from which the acquisition phase has to be carried forward.

- Event correlation: A huge number of events exist in the traditional digital environment itself and the scenario becomes intricate in the cloud environment. To handle the complexities, we add the phase of *event correlation* as a prerequisite for effective reconstruction. This phase identifies the associations among the events and groups them accordingly. The correlation phase should also relate the events pertaining to the target VM which are spread across multiple zones.
- Event reconstruction: The correlated events are given as input to the event reconstruction phase. Then the reconstruction of events can be done either solely based on the timeline or not completely on timeline based.

Taking all these phases as base, we devised an algorithm for cloud event reconstruction which we describe in the following subsection. The remaining common phases like notification, presentation will reflect the same functionality as in traditional digital environment. For example, during the presentation phase, validating the chain of custody is common in both the cloud and traditional digital environment.

3.2 Algorithm for Cloud Event Reconstruction

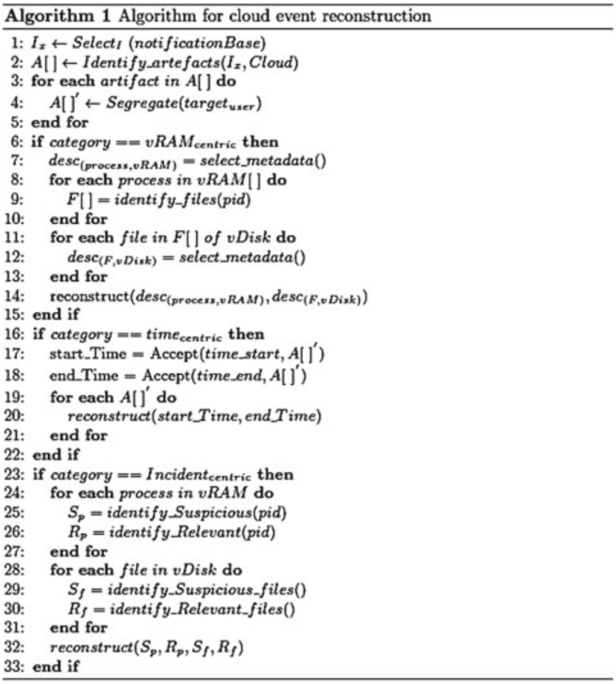
We propose an approach for reconstructing the events from multiple artifacts (Algorithm 1). The cloud provider takes the incident (I) reported and its details from the notification base. Then corresponding artifacts are identified (A[ ]) based on the incident (I_x) and the underlying cloud environment. Each identified artefact is filtered to get the events pertaining to the target user $(A[\ ]')$.

The inclusion of correlation phase for the process of event reconstruction gives the following advantages: (a) The number of events that are being considered for analysis is reduced (b) Interpretation of the events will be enhanced (c) The time spent on the investigation will be reduced. But, how the correlation can be performed and what are the correlation parameters are the important questions that one has to consider. In this context, we suggest the correlation process for the cloud environment which includes three major correlation parameters, i.e.,

- vRAM centric: In this category, we consider all the data present in the vRAM as base and correlate with the events in vDisk, i.e., initially, we retrieved all the files opened by each process in vRAM of the target virtual machine. Then, for each file identified in the vRAM, we extracted all its metadata from the target vDisk. There may be multiple ways to correlate the events in both the artifacts. But we have chosen file-based correlation as we observed various advantages like the number of events after correlation will be moderate, will aid the investigator to have faster interpretation, and will work irrespective of the VM operating system.

- Incident centric: Each incident can be detected with certain symptoms or characteristics. We can capture all those from the target virtual machine vRAM (S_p) and vDisk (S_f). In some cases, direct mapping of incident to symptoms may not be available due to the intruder anti-forensic measures. Then, the investigator has to look at the relevant events which may have an indirect association with the incident from both the artifacts i.e. from vDisk (R_f) and vRAM (R_p).
- Time centric: This sort of time-based correlation is very basic and can be helpful to the investigator when the above two are not functioning as expected. We accept the start time and end time in association with the target artifact. Then, using the same time range as reference, we search for the events present in artifacts of $A[\ ]'$.

The resultant events of each correlation category are given to the reconstruction function from which the timeline is generated. We validate the proposed algorithm in Sect. 4.



Algorithm 1 Algorithm for cloud event reconstruction

```
$I_x \leftarrow Select_I\ (notificationBase)$
$A[\,] \leftarrow Identify\_artefacts(I_x, Cloud)$
for each $artifact\ in\ A[\,]$ do
    $A[\,]' \leftarrow Segregate(target_{user})$
end for
if $category == vRAM_{centric}$ then
    $desc_{(process,vRAM)} = select\_metadata()$
    for each $process\ in\ vRAM[\,]$ do
        $F[\,] = identify\_files(pid)$
    end for
    for each $file\ in\ F[\,]\ of\ vDisk$ do
        $desc_{(F,vDisk)} = select\_metadata()$
    end for
    reconstruct$(desc_{(process,vRAM)}, desc_{(F,vDisk)})$
end if
if $category == time_{centric}$ then
    start_Time = Accept$(time\_start, A[\,]')$
    end_Time = Accept$(time\_end, A[\,]')$
    for each $A[\,]'$ do
        $reconstruct(start\_Time, end\_Time)$
    end for
end if
if $category == Incident_{centric}$ then
    for each $process\ in\ vRAM$ do
        $S_p = identify\_Suspicious(pid)$
        $R_p = identify\_Relevant(pid)$
    end for
    for each $file\ in\ vDisk$ do
        $S_f = identify\_Suspicious\_files()$
        $R_f = identify\_Relevant\_files()$
    end for
    $reconstruct(S_p, R_p, S_f, R_f)$
end if
```

4 Experiments and Results

4.1 Experimental Setup

We had setup a high end openstack private cloud with multiple nodes. The cloud had various tenants with each user having multiple virtual machines. The version of openstack used is Icehouse. All the major cloud services were running during our testing phase. We consider three major artifacts to perform event reconstruction, i.e., Service logs, vRAM, and vDisk.

4.2 Validation of the Proposed Framework—CERf

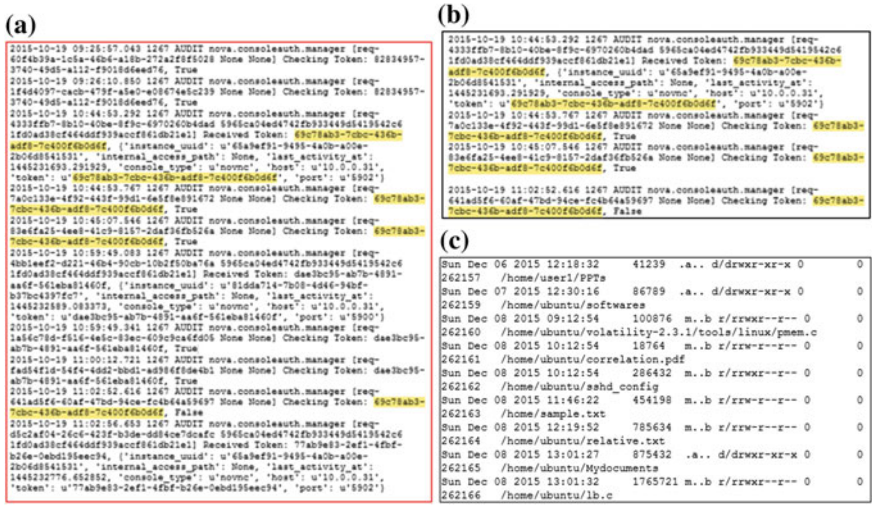
Segregation phase: As said, to preserve the privacy of other tenants, the process of segregation is required on each cloud artifact that was considered. But how this can be accomplished practically is challenging and tricky. We addressed this issue independently on each cloud artifact, i.e.,

- For service logs: Our core segregation parameter is "*instance_uuid*", i.e., we identified the target VM events based on this. But the problem with this approach is all the events in the service log(s) will not have *instance_uuid*. This makes the segregation process incomplete. We addressed this by identifying complementary parameters like *token id, request id, VM_IP* to enhance the overall segregation process. All these arguments can be validated with a simple example, i.e., there are many services (nova, image, networking, block storage service, etc.) that runs on Openstack cloud nodes and for each service there are different sets of logs. In Fig. 2a, a service log named *nova-consoleauth.log* was considered from the compute node of Openstack. It contains events of all the cloud tenants. We initially identified target VM events in the log with the instance_uuid and remaining events belonging to the same target are identified with another parameter-token id. A screenshot of the same is shown in Fig. 2b. Similar segregation process can be followed for any service log.
- Every tenant can have multiple users and in that case a single virtual machine may be used by more than one user. In such a situation, the process of segregation has to be applied on other artifacts as well like vRAM and vDisk. It is important to note that, the segregation parameters change from one artifact to the other. To make the illustration simpler, we consider a target tenant who is having one user so that explicit segregation is not required.

Event correlation phase: We captured all the events of Ubuntu 14.04 cloud virtual machine and noticed that there are thousands of events and it is difficult for the investigator to understand all of those. To enhance the event interpretation, we initially generated the *basic timeline* for the disk events of the victim VM (Fig. 2c).

Observations: (a) We observed from *basic timeline* in Fig. 2c that the number of events is still very high. (b) All the attributes shown by the timeline may not be useful for the investigator. (c) Representation of the simple timeline events shown in Fig. 2c is not sufficient and requires an advanced approach. (d) In some cases, the investigator may require some additional information which cannot be retrieved from the basic information presented in the timeline.

To address the above complications, we can apply the correlation process mentioned in the Algorithm 1. To make the illustration simple, we will show only the



(a)

```
2015-10-19 09:25:57.043 1267 AUDIT nova.consoleauth.manager [req-
60f4b39a-1c5a-46b6-a18b-272a2f8f5028 None None] Checking Token: 82834957-
3740-49d5-a112-f9018d6eed76, True
2015-10-19 09:26:10.850 1267 AUDIT nova.consoleauth.manager [req-
1f4d4097-cacb-479f-a5e0-e08674e5c239 None None] Checking Token: 82834957-
3740-49d5-a112-f9018d6eed76, True
2015-10-19 10:44:53.292 1267 AUDIT nova.consoleauth.manager [req-
4333ffb7-8b10-40be-8f9c-6970260b4dad 5965ca04ed4742fb933449d5419542c6
1fd0ad38cf464ddf939accf861db21e1] Received Token: 69c78ab3-7cbc-436b-
adf8-7c400f6b0d6f, {'instance_uuid': u'65a9ef91-9495-4a0b-a00e-
2b06d8541531', 'internal_access_path': None, 'last_activity_at':
1445231693.291929, 'console_type': u'novnc', 'host': u'10.0.0.31',
'token': u'69c78ab3-7cbc-436b-adf8-7c400f6b0d6f', 'port': u'5902'}
2015-10-19 10:44:53.767 1267 AUDIT nova.consoleauth.manager [req-
7a0c133e-4f92-443f-99d1-6e5f8e891672 None None] Checking Token: 69c78ab3-
7cbc-436b-adf8-7c400f6b0d6f, True
2015-10-19 10:45:07.546 1267 AUDIT nova.consoleauth.manager [req-
83e6fa25-4ee8-41c9-8157-2daf36fb526a None None] Checking Token: 69c78ab3-
7cbc-436b-adf8-7c400f6b0d6f, True
2015-10-19 10:59:49.083 1267 AUDIT nova.consoleauth.manager [req-
4bb1eef2-d221-46b4-90cb-10b2f50ba76a 5965ca04ed4742fb933449d5419542c6
1fd0ad38cf464ddf939accf861db21e1] Received Token: dae3bc95-ab7b-4891-
aa6f-561eba81460f, {'instance_uuid': u'81dda714-7b08-4d46-94bf-
b37bc4397fc7', 'internal_access_path': None, 'last_activity_at':
1445232589.083373, 'console_type': u'novnc', 'host': u'10.0.0.31',
'token': u'dae3bc95-ab7b-4891-aa6f-561eba81460f', 'port': u'5900'}
2015-10-19 10:59:49.341 1267 AUDIT nova.consoleauth.manager [req-
1a56c78d-f516-4e5c-83ec-609c9ca6fd05 None None] Checking Token: dae3bc95-
ab7b-4891-aa6f-561eba81460f, True
2015-10-19 11:00:12.721 1267 AUDIT nova.consoleauth.manager [req-
fad54f1d-54f4-4dd2-bbd1-ad986f8de4b1 None None] Checking Token: dae3bc95-
ab7b-4891-aa6f-561eba81460f, True
2015-10-19 11:02:52.616 1267 AUDIT nova.consoleauth.manager [req-
641ad5f6-60af-47bd-94ce-fc4b64a59697 None None] Checking Token: 69c78ab3-
7cbc-436b-adf8-7c400f6b0d6f, False
2015-10-19 11:02:56.653 1267 AUDIT nova.consoleauth.manager [req-
d5c2af04-26c6-423f-b3de-dd84ce7dcafc 5965ca04ed4742fb933449d5419542c6
1fd0ad38cf464ddf939accf861db21e1] Received Token: 77ab9e83-2ef1-4fbf-
b26e-0ebd195eec94, {'instance_uuid': u'65a9ef91-9495-4a0b-a00e-
2b06d8541531', 'internal_access_path': None, 'last_activity_at':
1445232776.652852, 'console_type': u'novnc', 'host': u'10.0.0.31',
'token': u'77ab9e83-2ef1-4fbf-b26e-0ebd195eec94', 'port': u'5902'}
```

(b)

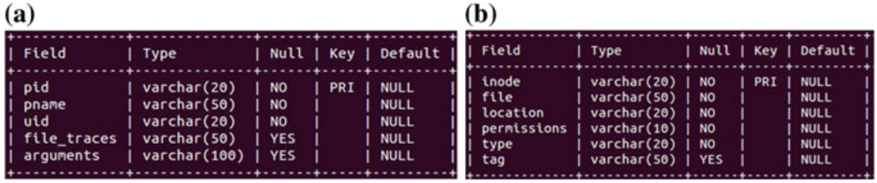
```
2015-10-19 10:44:53.292 1267 AUDIT nova.consoleauth.manager [req-
4333ffb7-8b10-40be-8f9c-6970260b4dad 5965ca04ed4742fb933449d5419542c6
1fd0ad38cf464ddf939accf861db21e1] Received Token: 69c78ab3-7cbc-436b-
adf8-7c400f6b0d6f, {'instance_uuid': u'65a9ef91-9495-4a0b-a00e-
2b06d8541531', 'internal_access_path': None, 'last_activity_at':
1445231693.291929, 'console_type': u'novnc', 'host': u'10.0.0.31',
'token': u'69c78ab3-7cbc-436b-adf8-7c400f6b0d6f', 'port': u'5902'}
2015-10-19 10:44:53.767 1267 AUDIT nova.consoleauth.manager [req-
7a0c133e-4f92-443f-99d1-6e5f8e891672 None None] Checking Token: 69c78ab3-
7cbc-436b-adf8-7c400f6b0d6f, True
2015-10-19 10:45:07.546 1267 AUDIT nova.consoleauth.manager [req-
83e6fa25-4ee8-41c9-8157-2daf36fb526a None None] Checking Token: 69c78ab3-
7cbc-436b-adf8-7c400f6b0d6f, True

2015-10-19 11:02:52.616 1267 AUDIT nova.consoleauth.manager [req-
641ad5f6-60af-47bd-94ce-fc4b64a59697 None None] Checking Token: 69c78ab3-
7cbc-436b-adf8-7c400f6b0d6f, False
```

(c)

```
Sun Dec 06 2015 12:18:32     41239  .a.. d/drwxr-xr-x 0         0
262157   /home/user1/PPTs
Sun Dec 07 2015 12:30:16     86789  .a.. d/drwxr-xr-x 0         0
262159   /home/ubuntu/softwares
Sun Dec 08 2015 09:12:54     100876  m..b r/rrwxr--r-- 0        0
262160   /home/ubuntu/volatility-2.3.1/tools/linux/pmem.c
Sun Dec 08 2015 10:12:54     18764   m..b r/rrw-r--r-- 0        0
262161   /home/ubuntu/correlation.pdf
Sun Dec 08 2015 10:12:54     286432  m..b r/rrwxr--r-- 0        0
262162   /home/ubuntu/sshd_config
Sun Dec 08 2015 11:46:22     454198  m..b r/rrw-r--r-- 0        0
262163   /home/sample.txt
Sun Dec 08 2015 12:19:52     785634  m..b r/rrw-r--r-- 0        0
262164   /home/ubuntu/relative.txt
Sun Dec 08 2015 13:01:27     875432  .a.. d/drwxr-xr-x 0        0
262165   /home/ubuntu/Mydocuments
Sun Dec 08 2015 13:01:32     1765721 m..b r/rrwxr--r-- 0        0
262166   /home/ubuntu/lb.c
```

Fig. 2 **a** Openstack service log-nova-consoleauth before segregation. **b** Nova-consoleauth after segregation with our identified parameters. **c** A simple timeline file without any filtering



(a)

```
+-------------+--------------+------+-----+---------+
| Field       | Type         | Null | Key | Default |
+-------------+--------------+------+-----+---------+
| pid         | varchar(20)  | NO   | PRI | NULL    |
| pname       | varchar(50)  | NO   |     | NULL    |
| uid         | varchar(20)  | NO   |     | NULL    |
| file_traces | varchar(50)  | YES  |     | NULL    |
| arguments   | varchar(100) | YES  |     | NULL    |
+-------------+--------------+------+-----+---------+
```

(b)

```
+-------------+-------------+------+-----+---------+
| Field       | Type        | Null | Key | Default |
+-------------+-------------+------+-----+---------+
| inode       | varchar(20) | NO   | PRI | NULL    |
| file        | varchar(50) | NO   |     | NULL    |
| location    | varchar(20) | NO   |     | NULL    |
| permissions | varchar(10) | NO   |     | NULL    |
| type        | varchar(20) | NO   |     | NULL    |
| tag         | varchar(50) | YES  |     | NULL    |
+-------------+-------------+------+-----+---------+
```

Fig. 3 **a** Normalized schema for target VM-vRAM. **b** Target VM vDisk normalized schema

VRAM centric correlation, i.e., to identify the files in vDisk which had traces in vRAM. We used *volatility*, a memory analysis tool to achieve this. To get a comprehensive picture of the entire virtual machine, the timeline should contain events from other artifacts as well. But this involves various technical challenges: (a) Events from multiple artifacts will have different formats and representing those events in a single timeline is difficult. This requires all those events to convert to a unique format. (b) This normalization process has to be done carefully otherwise the chance of data loss will be high. Taking these in to consideration, the following process is suggested.

Normalization process for vRAM: There is a lot of metadata associated with each process in vRAM. Considering all those attributes for timeline will overload the information provided by each event. So we identified more significant attributes for vRAM events. But these attributes were spread across multiple plugins of *volatility*. Then we parsed each plugin output and stored in MySQL database. Later, the significant attributes from different plugins were joined to form a new relational table and it has the schema as shown in the Fig. 3a.

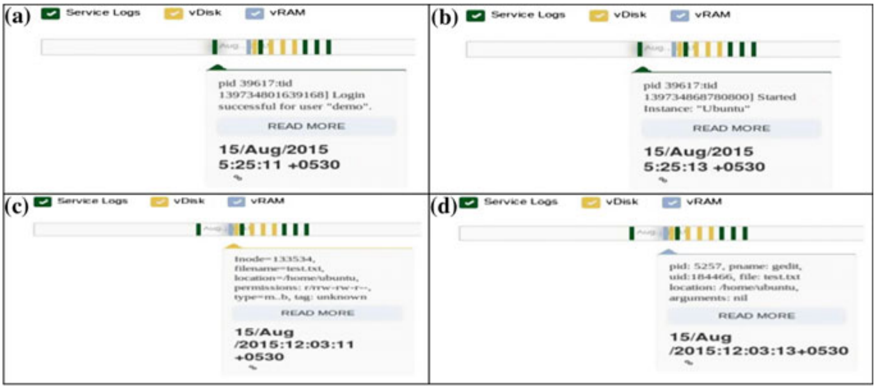
Normalization process for vDisk: We consider only the disk events which had traces in vRAM. But each event in disk will have lot of associated attributes. We considered only the important attributes and represented in our advanced visual timeline and the corresponding schema is shown in Fig. 3b.

The significant attributes identified from both the artifacts are normalized to a format <*Time*, *Description*>. Here, the *description* is the main attribute which contains sub-attributes shown in Fig. 3a, b. It is important to note that, care has been taken to inject only the sub-attributes with high importance to the final normalized schema. The significance of each sub-attribute is decided based on two properties: (1) The sub-attribute should be present in most of the event categories (2) With each significant sub-attribute an event should be recognized uniquely or at least should be able to extract other attributes data which is not explicitly considered for timeline generation.

Then, we identified the processes in vRAM which are using files from the vDisk. We then filtered, extracted, and considered only those files and the associated significant sub-attributes from the vDisk to generate the timeline. Then the number of events in Fig. 2c came down and this benefits the investigator to complete the forensic process faster.

Event reconstruction phase: The reduced set of correlated events is given as input to the reconstruction phase. But this introduces new challenge for the investigator, i.e., regarding the comprehensiveness involved in analyzing the timeline with some events being filtered. We handle this in two stages:

- Stage 1: Add events to the timeline which should describe at least the below events: (a) *VM creation time* (b) *VM logout time* (c) *System IP from which cloud was accessed* (d) *Highly critical events happened in the victim VM* (e) *VM termination time* (f) *Suspicious events in the target VM* etc.,
  We identified that the following information is present in the artifact *service logs*. Including the above information in the timeline will make the investigator to know the exact start time (VM creation/booting/login time etc.,) from which the process of forensic examination can be easily initiated. For example, in Fig. 4a, the target user login time is very obvious. Moreover, in Fig. 4b we also included the next immediate activity after login, i.e., *started an instance named ubuntu*. Then that VM events are tracked and reconstructed using the timeline. Due to the space restriction, we cannot present each timeline event. But the major ones are present in Fig. 4c—an event from target VM disk, Fig. 4d—an event from the target vRAM. Moreover, tracking and reconstructing the target VM events can be stopped when that VM is logged out or terminated (Fig. 5).
- Stage 2: In all the cases, the information provided by the timeline may not be sufficient for the investigator. We handled those cases by providing additional attribute information associated with each event of every artifact type. Some of the other important attribute information which we consider is shown in Table 1. This information is presented to the user when he interacts only with *Read More* section of the corresponding event in the timeline.



(a)

pid 39617:tid 139734801639168] Login successful for user "demo".

15/Aug/2015 5:25:11 +0530

(b)

pid 39617:tid 139734868780800] Started Instance: "Ubuntu"

15/Aug/2015 5:25:13 +0530

(c)

Inode=133534, filename=test.txt, location=/home/ubuntu, permissions: r/rrw-rw-r--, type=m..b, tag: unknown

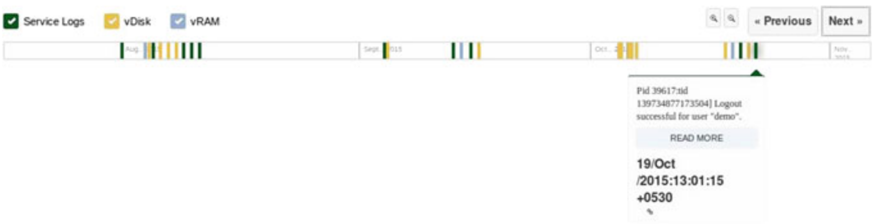
15/Aug /2015:12:03:11 +0530

(d)

pid: 5257, pname: gedit, uid:184466, file: test.txt location: /home/ubuntu, arguments: nil

15/Aug /2015:12:03:13+0530

Fig. 4 **a** Timeline event from service logs-login event. **b** An event from service logs-Instance start up time. **c** Timeline event from target vDisk-a file event. **d** Timeline event from target vRAM-a process event



Pid 39617:tid 139734877173504] Logout successful for user "demo".

19/Oct /2015:13:01:15 +0530

Fig. 5 Timeline event from service logs-instance logout event

The final timeline includes the events from victim VM vDisk, vRAM, and even from service logs. The advantages of this customized timeline are: (a) Interpretation of the events will be easier (b) Comprehensiveness of the data in the timeline is more as it includes events from multiple artifacts (c) We included highly related and common events in the final timeline and this makes the investigation to complete faster (d) Flexibility to look at the data of other related events is made possible with our customized timeline.

4.3 *Other Parameters Considered by Our CERf (Cloud-Based Event Reconstruction Framework)*

- Completeness Versus Utility: If all the events from the cloud artifacts are present in the timeline then it may not be useful for the investigator. On the other side, if we remove more number of events from the timeline then the comprehensiveness may not be achieved which may lead to misleading conclusions. So there should

Table 1 Additional attributes considered for each artifact

Other metadata considered for each artifact		
vRAM	vDisk	Service logs
Offset	Size	An event reflecting symptoms
UID	Hash value	Traces of every error event
GID	Sector range	Error based events
DTB	Hex value	Less critical events
Parent–child and network_connections	Tag parameter	Resource usage

be proper trade off between completeness and utility which we achieved by our framework CERf.

- Scalability: Since the events are filtered and presented, the complications involved in the issue of scaling may get reduced. Moreover, we applied filtering at multiple levels and considered events of target VM from multiple nodes.
- Extensibility: Since timeline-based reconstruction is applied to the massive cloud environment, it can be easily extended to other environments as well especially when they involve vRAM, vDisk and logs as artifacts. In [17], the authors used timeline-based event construction. They reduced the number of events by converting some of the low level events to high level events. But their approach is not extensible even to the different version of an operating system or to a different file system as they purely used rule-based approach for achieving this. In our approach, these issues will not arise.

5 Conclusion

The need for forensic standards and methodologies in cloud is increasing along with the growing popularity and usage of cloud. There are many technical challenges associated with the domain of cloud forensics. In this paper, we focus on one of the technical challenges namely *Event reconstruction*. We identified that event reconstruction can be performed by using timeline and without explicit timeline. We focused on the former approach. We proposed a framework called as Cloud-based Event Reconstruction framework (CERf). The framework is validated with the corresponding experiments. In this connection, we identified a way to generate a comprehensive advanced timeline which contains events from multiple cloud artifacts. Our approach of timeline-based event reconstruction benefits the investigator in various aspects like scalability, utility, and extensibility.

As a future work, we are going to add more filters at the correlation phase to further reduce the number of events for the reconstruction phase. Moreover, in this paper we tested our framework CERf in private cloud environment and planning to extend the same for the public cloud platforms.

References

1. Sutte J.: Twitter hack raises questions about cloud computing, In: http://edition.cnn.com/2009/TECH/07/16/twitter.hack/, (2009), accessed 21-07-2013.
2. Higgins K.: Dropbox, wordpress used as cloud cover in new apt attacks, In: http://www.darkreading.com/attacksbreaches/dropbox-wordpress-used-as-cloud-coverin/240158057, (2013), accessed 22-07-2013.
3. Inci, Mehmet Sinan, et al.: Seriously, get off my cloud! Cross-VM RSA Key Recovery in a Public Cloud. In: IACR Cryptology ePrint Archive, 2015.
4. NIST Cloud Computing Forensic Science Challenges, Tech. Rep. In: http://csrc.nist.gov/publications/drafts/nistir-8006/draft_nistir_8006.pdf.
5. Chabot, Yoan, et al.: Event Reconstruction: A State of the Art. In: Handbook of Research on Digital Crime, Cyberspace Security, and Information Assurance (2014): 15.
6. Cohen, Fred.: Challenges to Digital Forensic Evidence in the Cloud. In: Cybercrime and Cloud Forensics: Applications for Investigation Processes (2012): 59.
7. Ruan K., J. Carthy, T. Kechadi, M. Crosbie.: Cloud Forensics, In: 7th IFIP Advances in Digital Forensics VII, G. Peterson and S. Shenoi (eds), vol. 361, pp. 35–46.
8. Ruan K., Carthy, J.: Cloud Computing Reference Architecture and its Forensic Implications: a Preliminary Analysis, In: Proceedings of the 4th International Conference on Digital Forensics & Cyber Crime, Springer Lecture Notes, October 25–26, Lafayette, Indiana, USA.
9. James, Joshua I., Ahmed F. Shosha, and Pavel Gladyshev.: Digital forensic investigation and cloud computing. In: Cybercrime and Cloud Forensics: Applications for Investigation Processes (2012): pp. 1–41.
10. Spyridopoulos, Theodoros, and Vasilios Katos.: Data Recovery Strategies for Cloud Environments. In: Cybercrime and Cloud Forensics: Applications for Investigation Processes (2012): 251.
11. Garfinkel, Simson L.: Automating disk forensic processing with SleuthKit, XML and Python. In: Fourth International IEEE Workshop on Systematic Approaches to Digital Forensic Engineering, SADFE'09, IEEE, 2009.
12. Garber, Lee.: Encase: A case study in computer-forensic technology. In: IEEE Computer Magazine January (2001).
13. Olsson, Jens, and Martin Boldt.: Computer forensic timeline visualization tool. In: digital investigation, pp. S78–S87, Elsevier, 2009.
14. Kristinn.: Mastering the super timeline with log2timeline. In: SANS Institute (2010).
15. Buchholz, Florian P., and Courtney Falk.: Design and Implementation of Zeitline: a Forensic Timeline Editor. In: DFRWS. 2005.
16. Agarwal, Ritu, and Suvarna Kothari.: Review of Digital Forensic Investigation Frameworks. In: Information Science and Applications. Springer Berlin Heidelberg, pp. 561–571, 2015.
17. Hargreaves, Christopher, and Jonathan Patterson.: An automated timeline reconstruction approach for digital forensic investigations. In: Digital Investigation, pp. S69–S79, 2012.

eCloud: An Efficient Transmission Policy for Mobile Cloud Computing in Emergency Areas

Bibudhendu Pati, Joy Lal Sarkar, Chhabi Rani Panigrahi and Shibendu Debbarma

Abstract Due to the less resources and battery power of mobile devices or intermittent connectivity between mobile devices and cloud, the users may face huge difficulties. There are a very few work have been proposed which can solve these problems. But, to select best cloud for mobile devices as well as to minimize the transmission latency upon mobility of mobile nodes, in this work an efficient transmission policy for Mobile Cloud Computing (MCC) named as eCloud is proposed. In eCloud, mobile nodes can select their best cloud for sending requests upon mobility of mobile nodes. eCloud can solve the problem of battlefield situation or any emergency condition like earthquake or terrorists attack.

Keywords Mobile cloud computing · Offloading · Energy-efficiency

1 Introduction

The number of users of mobile devices is increasing day by day [1]. Although different users have different demands. For example, some of the users have interest in sports, gaming and some of the users are interested for other entertainments [2]. From research point of view, several users want to execute their code in MATLAB

B. Pati (✉) · J.L. Sarkar
Department of Computer Science and Engineering,
C.V. Raman College of Engineering, Bhubaneswar, India
e-mail: patibibudhendu@gmail.com

J.L. Sarkar
e-mail: joylalsarkar@gmail.com

C.R. Panigrahi
Department of Computer Science, Central University of Rajasthan,
Bandar Sindri, Rajasthan, India
e-mail: panigrahichhabi@curaj.ac.in

S. Debbarma
Department of Information Technology, Tripura University, Agartala, India
e-mail: shibendu@gmail.com

P.K. Sa et al. (eds.), *Progress in Intelligent Computing Techniques: Theory, Practice, and Applications*, Advances in Intelligent Systems and Computing 719, DOI 10.1007/978-981-10-3376-6_5

or other simulation software. Moreover, there are large number of users who use various social media in their mobile devices and perform different activities such as uploading their photos, videos etc. [3]. But, in each case users may face difficulties with their devices as mobile devices have limited battery power, less resources, and memory etc. [4, 5]. The popular cloud computing technology overcomes these problems and creates more interest to the users [6]. In this work, authors present an energy efficient transmission policy between mobile devices and cloud or cloudlets named as eCloud which is mainly developed based on the services as given in [4]. Where, mobile nodes can have simultaneous access to the several networks and can choose best cloud for sending their data requests or vice versa.

The rest of the paper is organized as follows: Sect. 2 presents related work. Section 3 describes the efficient transmission policy for eCloud. Section 4 presents mobility management for mobile node with failure and repair. Section 5 presents the results obtained along with the analysis of results. Finally, Sect. 6 concludes the paper.

2 Related Work

For reducing energy consumption of mobile devices by increasing the battery life, there are a few works have been done which support offloading scheme [1–3]. In [1], authors proposed a scheme to reduce the energy consumption of mobile devices which supports offloading scheme. In [2, 3] authors proved that by using the remote sensing, the battery consumption can be reduced for different tasks. In [7], authors investigated a method called Thinkair which provides a method-level computation offloading to the cloud. Thinkair mainly works on the offloading scheme from mobile devices to the cloud and also the cloud scalability. In [8], authors proposed a context-aware decision engine which supports offloading system from mobile devices to cloud where to make the offloading decision. The proposed method does not support multiple cloud resources and the availability of the wireless channel and information of the geographical location is incorrect when a device is indoor.

3 An Efficient Transmission Policy

Due to high user rate of mobile devices, every mobile device should be energy efficient. Because different users have different demands. Each mobile device can offload its application to the Local Cloud (LC). Mobile devices connect with LC for reducing latency as well as storage. The LC can also connects with the Command Station (CS). The CS can handle LC as well as mobile devices in case of any emergency occurs mainly earth quake, terrorist attack, etc. In case of handoff mechanism, mobile devices can connect with the LC by WiFi, 3G, or by other networks. Due to the signal fading or other reason if QoS of WiFi does not fulfill the requirements, the

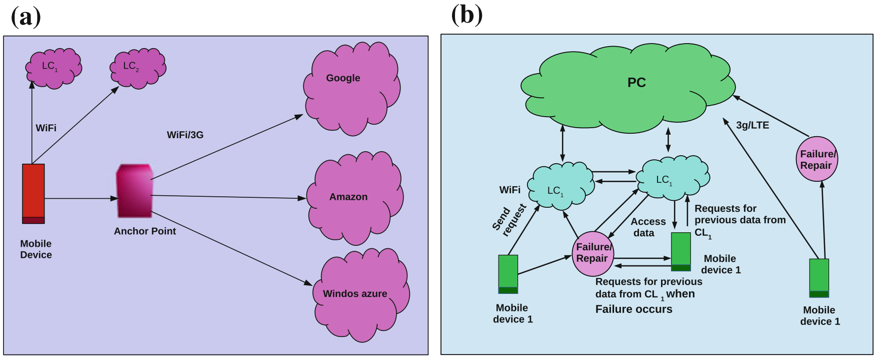


Fig. 1 **a** eCloud: using AP, LC, and PC. **b** Mobility management of eCloud

mobile device can connect to LC by 3G. Each mobile device also can connect with the Public Cloud (PC) when failure occurs. One of the main advantage of eCloud is that mobile devices can connect to several access points simultaneously before initializing handoff mechanism. eCloud managed Multi-homed Mobile IP (M-MIP) for efficient transmission between mobile devices and LC. An Anchor Point (AP) is used which can work as a home agent as shown in Fig. 1a. Each mobile device selects its best cloud by the method described in [4]. The Network Probing Service (NPS) is maintained by calculating the Relative Network Load (RNL) of each network step. An AP runs the Cloud Probing Service (CPS) and Cloud Ranking Service (CRS) for probing between LC and PC and chooses best cloud for mobile devices and mobile devices then offload their applications to the LC or PC.

Algorithm 1: eCloud

Step 1: Initially for each time slot t_i and for each application a_i monitoring real-time bandwidth between mobile device and LC
Step 2: Each mobile device maintain M-MIP for simultaneous access with several access points
Step 3: AP runs the CPS and CRS
Step 4: Choose best PC or LC
Step 5: Each mobile device sends data request to the LC
Step 6: If mobile device does not access data from the previous LC, it then sends request to the current LC
Step 7: Mobile devices access previous data from the current LC
Step 8: If their is no LC in the range of mobile devices then they connect with the PC

In Algorithm 1, initially eCloud determines time slot t_i for each application of mobile devices for monitoring real-time bandwidth and also for simultaneous access with several access points mobile devices maintain M-MIP (Steps 1–2). An AP runs on the CRS and on CPS by which a mobile device selects its best cloud and also probes between LC and PC (Steps 3–4). Mobile devices send data request to the LC and if a device does not receive any reply from this LC, the mobile device then sends a request to the current LC and accesses those data from the current LC (Steps 5–7). Due to mobility, mobile devices shift from one place to another and if there is no LC within its range then the devices connect with the PC.

4 Mobility Management for Mobile Node with Failure and Repair

The mobility management of mobile nodes in eCloud is shown in Fig. 1b. Where, each mobile device sends data request to the LC. During transmission of data, there maybe failure occurs due to the unavailability of network, mobility, etc., and mobile devices can send data request to the LC after repairing those failures. If mobile devices do not have access data from current LC, where a mobile device is in the range of another LC then the mobile device sends request to the current LC for accessing data from the previous LC. The mobile devices then access those data from the current LC. During transition from one place to another if there are no LCs present in the range of mobile devices, the mobile devices then connect with the PC.

Let us assume that $m_1, m_2, m_3..., m_n \in M$ be the total number of mobile devices and then the total number of mobile nodes will be as given in Eq. (1):

$$t_d = (m_1 + m_2 + m_3 + ... + m_n) = \sum_{i=1}^{n} m_i. \tag{1}$$

Now, at time t_i seconds, let k number of mobile devices denoted as $\check{m}_1, \check{m}_2, ..., \check{m}_k$ shift to another region. So, after time t_i seconds the number of mobile devices present at earlier region will be as given in Eq. (2):

$$t_d = \sum_{i=1}^{n} m_i - (\check{m}_1 + \check{m}_2 + \check{m}_3 + ... + \check{m}_k) = \sum_{i=1}^{n} m_i - \sum_{i=1}^{k} \check{m}_i. \tag{2}$$

Let, each device has different number of tasks and is given as in Eq. (3):

$$\kappa_T^l = \sum_{i=1}^{a} T_i + \sum_{j=1}^{b} T_j + ... + \sum_{k=1}^{c} T_k \tag{3}$$

Here, there are b, c,..., m be the total number of tasks of $m_1, m_2, m_3..., m_k$, respectively. Let $\alpha, \beta, ...\gamma$ be the amount of tasks that are reduced from the mobile devices due to offloading then the total number of tasks after offloading denoted as κ_T^r and is given in Eq. (4):

$$\kappa_T^r = \sum_{i=1}^{a} T_i - \alpha + \sum_{j=1}^{b} T_j - \beta + ... + \sum_{k=1}^{c} T_k - \gamma. \tag{4}$$

Now, each $m_1, m_2, ..., m_k \in M$ can connect with LC in that region. But, due to the mobility of mobile nodes, the connection maybe failed after some time. Let, ϕ is the period of time for recovering (repair) upon failure occurs and Y be the failure rate. So the expectation time of Failure/Repair state can be computed using equation as given in [4] and can be represented as in Eq. (5).

$$t(\phi^*) = \frac{1}{Y}\left(\frac{1}{t[e^{-Y.\phi}]} - 1\right)\left(\sum_{i=1}^{n} m_i - \sum_{i=1}^{k} \check{m}\right) \tag{5}$$

Let us consider, mobile nodes and LC are placed in a 2D-space and is given as in Eqs. (6) and (7):

$$A_d = \int_x \int_y \delta x \delta y, \tag{6}$$

$$A_l = \int_m \int_n \delta u \delta v. \tag{7}$$

So, there should be region where mobile nodes and LCs are closed to each other and in such region the overall failure can be minimized due to the mobility of mobile nodes and is given in Eqs. (8) and (9):

$$R_d^l(min) = \int_x \int_y \delta x \delta y \cap \int_m \int_n \delta u \delta v, \tag{8}$$

$$\int_x \int_y \delta x \delta y \cap \int_m \int_n \delta u \delta v \approx min\left(\frac{1}{Y}\left(\frac{1}{t[e^{-Y.\phi}]} - 1\right)\left(\sum_{i=1}^{n} m_i - \sum_{i=1}^{k} \check{m}\right)\right). \tag{9}$$

5 Simulation

The proposed approach eCloud was evaluated based on Android operating system. Android x86 was installed on Intel I3 laptop. Samsung 1997 was used for deploying the applications. eCloud was run in a standalone environment, where unwanted applications were completely closed and background jobs were also shutdown. The static and dynamic power of CPU was set from 0.3 to 1 randomly and also the clock frequency was set ranging from 1.2 GHz to 1.6 GHz randomly. For LC, a Dell I3 computer was used with 8 GB RAM and for PC, Dell I7 computers with a RAM of 16 GB, and 2.8 GHz processor was used. An Activity Recognition (AR) web service is running on the mobile nodes which is built by Java and after running AR, it

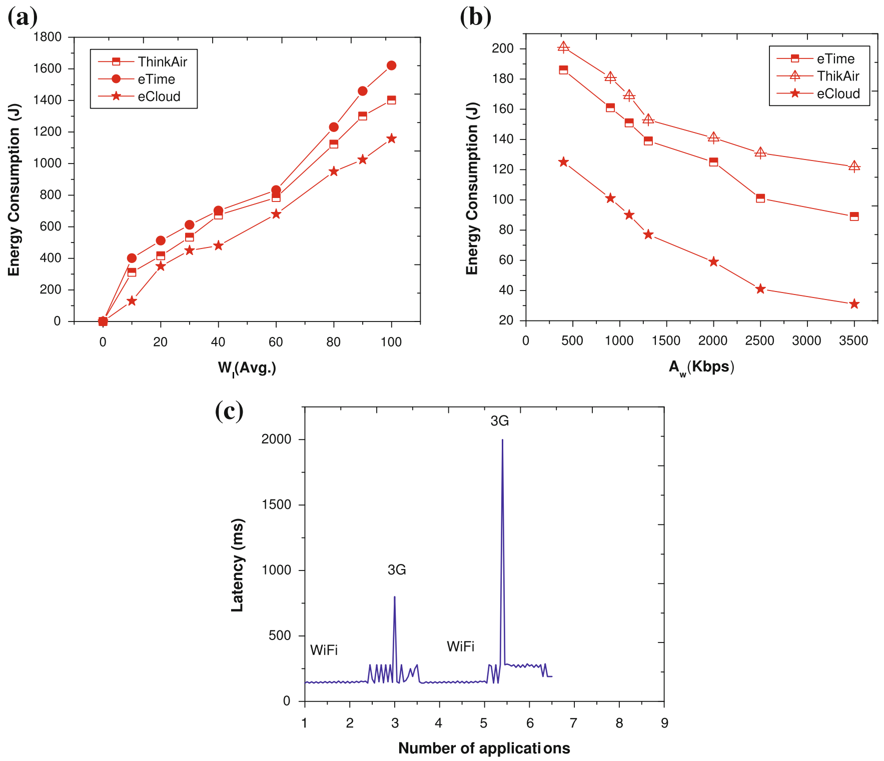


Fig. 2 **a** Total energy consumption of *eCloud* under different workloads of eCloud with two different baselines. **b** Comparison of energy consumption for downloading 15 MB application of *eCloud* with two baselines using WiFi. **c** Latency for receiving data under different applications using WiFi, 3G, and bluetooth services

then registered with both LC and PC. Both CRS and CPS were running on AP and M-MIP protocol and were developed in Java environment.

5.1 Results and Analysis

Figure 2a shows the comparison between *eCloud* with two different baselines called *eTime* and *Thinkair*. Figure 2a shows that energy consumption is low in case of *eCloud* when average workload (w_l) varied. Next, the workload is changed by increasing the number of running applications and then the average is taken. Figure 2b shows the energy consumption of eCloud with two different baselines, where different bandwidths (A_w) (in Kbps) were considered. The energy consumption was found to be less in *eCloud* with respect to two baselines. Figure 2c shows the overall latency when AR is running on the LC with different number of applications. From Fig. 2c, it is clear that the latency is high for 3G and bluetooth services as compared to WiFi.

6 Conclusion

In this work, an efficient approach for MCC is proposed which can be useful in emergency condition and is named as eCloud. eCloud chooses best cloud for mobile nodes and nodes can have simultaneous access with different access point by M-MIP. During mobility of mobile nodes in eCloud, the mobile nodes can send requests to the nearest LC for accessing data from the previous LC.

References

1. Panigrahi, C. R., Pati, B., Tiwary, M., Sarkar, J. L.: EEOA: Improving energy efficiency of mobile cloudlets using efficient Offloading Approach. IEEE International Conference on Advanced networks and telecommunications systems, pp. 1–6, (2015).
2. Rudenko, A., Reiher, G. P. P., and Kuenning, G.: Saving portable computer battery power through remote process execution. In Mobile Computing and Commun. Review, **2**, pp. 19–26 (1998).
3. Rudenko, A., Reiher, G. P. P., and Kuenning, G.: The remote processing framework for portable computer power saving. In Proc. 1999 ACM Symposium on Applied Computing, pp. 365–372 (1999).
4. Mitra, K., Saguna, Åhlund, C.: A Mobile Cloud Computing System for Emergency Management. IEEE Cloud Computing, **43**(4), pp. 30–38 (2014).
5. Bowen, Z., Dastjerdi, A. V., Calheiros, R. N., Srirama, S. N., and Buyya, R.: A Context Sensitive Offloading Scheme for Mobile Cloud Computing Service. In Proceedings of the IEEE 8th International Conference on Cloud Computing, pp. 869–876 (2015).
6. Rong, P. and Pedram, M.: Extending the lifetime of a network of battery powered mobile devices by remote processing: a Markovian decision based approach. In Proc. 2003 Annual Design Automation Conference, pp. 906–911 (2013).
7. Kosta, S., Aucinas, A., Hui, P., Mortier, R., and Zhang, X.: Thinkair: Dynamic resource allocation and parallel execution in the cloud for mobile code offloading. In Proceedings of 31st IEEE International Conference on Computer Communications, pp. 945–953 (2012).
8. Shu, P., Liu, F., Jin, H., Chen, M., Wen, F., Qu, Y., Li, B.: eTime: Energy-Efficient Transmission between Cloud and Mobile Devices. IEEE Infocom, pp. 195–199 (2013).

A Multidimensional Approach to Blog Mining

K.S. Sandeep and Nagamma Patil

Abstract Blogs are textual web documents published by bloggers to share their experience or opinion about a particular topic(s). These blogs are frequently retrieved by the readers who are in need of such information. Existing techniques for text mining and web document mining can be applied to blogs to ease the blog retrieval. But these existing techniques consider only the content of the blogs or tags associated with them for mining topics from these blogs. This paper proposes a Multidimensional Approach to Blog Mining which defines a method to combine the Blog Content and Blog Tags to obtain Blog Patterns. These Blog Patterns represent a blog better when compared to Blog Content Patterns or Blog Tag Patterns. These Blog Patterns can either be used for Blog Clustering or used by Blog Retrieval Engines to compare with user queries. The proposed approach has been implemented and evaluated on real-world blog data.

Keywords Blogs · Blog mining · Tags · Blog clustering

1 Introduction

Blogs are web documents written about a particular topic(s). These topics maybe related to current events of the world, technology, gadgets, or anything else. Bloggers explain their understanding or opinion about these topics in their blogs. So, it can be said that these blogs are the source of huge amount of information. With the widespread use of Internet, there is a phenomenal increase in the number of bloggers and readers who follow them. The idea expressed in the blogs can influence the readers and hence can be decisive. So, there is a need to extract the information

K.S. Sandeep (✉) · N. Patil
Department of Information Technology, National Institute of Technology Karnataka,
Surathkal, Mangalore 575025, Karnataka, India
e-mail: sandy89rao89@gmail.com

N. Patil
e-mail: nagammapatil@nitk.ac.in

P.K. Sa et al. (eds.), *Progress in Intelligent Computing Techniques: Theory, Practice, and Applications*, Advances in Intelligent Systems
and Computing 719, DOI 10.1007/978-981-10-3376-6_6

present in the blogs. This leads us to Blog Mining, which refers to information mining from the blogs.

Along with the text content, blogs also contain other information such as title of the blog, author of the blog, date and time when the blog was published, tags or categories attribute and others. Each of these can be considered as a dimension of the blog. Out of these dimensions, blog content and blog tags are the most interesting ones which are significant for the information extraction process. Blog Content is the actual write up by the blogger which is in form of text. It is possible to extract the information from the blog content by using Text Mining techniques. But the extracted information may also contain noise due to the way blogs are written. Blog Tags refer to the topics on which the blog is written and they are specified either by bloggers and readers. It is possible to cluster the blogs based on these tags information. But the tags may not be consistent as they are specified by bloggers and readers. Both of these dimensions are useful but have their own drawbacks. So, this paper proposes a method to combine the results of Blog Content Mining and Blog Tag Mining to obtain Blog Patterns which represent the blog in a better way when compared to Blog Content Patterns and Blog Tag Patterns.

The rest of the paper is organized as follows: Section 2 provides an overview of related work in the field of blog mining and text mining. Section 3 explains the proposed methodology for combining Content Mining and Tag Mining approaches. Section 4 explains the procedure used to evaluate the proposed method. Section 5 provides the results and its analysis. Section 6 provides the conclusion.

2 Related Work

Many techniques have been proposed for mining the text documents and these techniques can be applied to blogs to perform Blog Content Mining. There are also notable works in the field of Social Tag Mining in general and Blog Tag Mining in particular.

Daniel EOLeary [1] has given a detailed explanation about the blogs, Blog Mining, and Sentiment Analysis from blogs. First, it explains the importance of blogs by quoting Polanyi's work [2] about two types of knowledge: explicit and tacit. Explicit Knowledge is more of facts and is easily communicated and documented. But, Tacit Knowledge is the one gained with experience and cannot be easily documented, but it is valuable. The author claims that bloggers document such tacit knowledge in their blogs and hence there is a need to extract such information from them. It also explains how important blogs are for product/service based organizations where the organization can get effective feedback about their product/service from these blogs. Next, this work explains in some detail the different ways to select a sample of blogs for analysis and limitations in each way. Then the work moves to explain Sentiment Analysis in general by using Opinion Word Dictionaries. The same technique is applied to financial blogs to determine stocks with negative opinions.

Zhang et al. [3] proposed a technique for blog clustering by exploiting the relation between tags and blogs. It states that each blog maybe associated with a set of tags and each tag maybe associated with a set of blogs. So, the work aims to cluster all the blogs having similar tags together using the AGNES Clustering algorithm. Flora S Tsai [4] proposed a tag-topic model for blog mining. The work mainly explains the multidimensional nature of the blogs. It argues that Blog Content Mining has its own limitations since the blog text may include abbreviations, slang words, spelling and grammatical errors, and in the worst case may also include terms of different languages. On top of this, processing the complete text content of entire blog collection is time consuming. So it proposes blog mining based on tags. Even though the tags can be inconsistent, the work claims that it is possible to reduce the noise using certain probabilistic and dimensionality reduction techniques. It applies the Latent Dirichlet Allocation method on the tags associated with the blogs to extract the appropriate tags for that blog. Yi-Hui Chen et al. [5] proposed a keyword-based method for Blog Mining by analyzing user behaviors. They note that extracting keywords or topics by mining each and every blog completely in the entire blog collection can be time-consuming. So the work proposes that whenever user provides a search text and a set of blogs are retrieved for that search text, it is possible to associate the keywords present in the search text with that blog. It uses the Full-text Keyword Retrieval Process based on Term Frequency and Inverse Document Frequency to extract the set of keywords from the blogs and also from search text. It compares both the keyword sets to examine the difference between the two and finally it combines the keywords sets to obtain a single set. Ning Jong et al. [6] proposed an effective pattern discovery approach for text mining based on pattern taxonomy model. This model considers each document as a set of paragraphs and each paragraph as a set of terms. Then it determines closed frequent patterns from the text document. All such closed frequent patterns are said to represent the text document in a way better than the frequent terms or frequent phrases do.

Yuefeng Li et al. [7] propose a relevance feature discovery model for text mining which determines both positive and negative patterns from a text document. Duc-Thuan Vo et al. [8] proposes a classification model for short text based on topic modeling where short text classification is challenging due to data sparseness. Xiang Wang et al. [9] propose a topic mining approach for text sequences based on timestamps. Chenghua Lin et al. [10] propose a joint sentiment-topic modeling technique based on Latent Dirichlet Allocation (LDA) method. Senghua Bao et al. [11] propose another approach for emotion topic modeling of affective text based on LDA.

All the above works provide different approaches that can be used for mining topics from text content and tags associated with blogs.

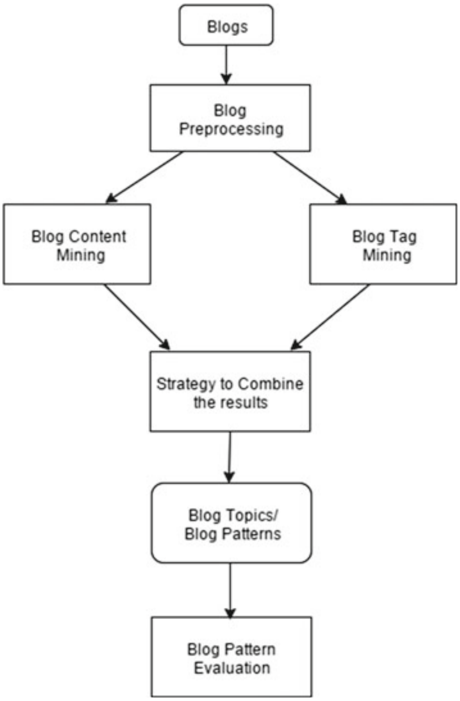


Fig. 1 Proposed methodology for blog mining

3 Proposed Methodology

The methodology proposed here is to determine the final topic patterns of a blog called as Blog Patterns by combining both, the Content Patterns derived from the text content of the blog and the Tags Patterns derived from tags associated with that blog. The block diagram of proposed approach is given in Fig. 1.

Stage 1: Preprocessing of blog text and blog tags is performed by applying Stopwords Removal and Stemming Techniques.

Stage 2: Pattern Taxonomy Model [6] is applied to the preprocessed text content to mine the set of closed frequent patterns. This model considers each document as a set of paragraphs and each paragraph as a set of terms. Term frequency is defined as the number of paragraphs in which the term occurs. Sequential Frequent Patterns of all lengths are obtained from this model and checked for closure. The set of all Closed Sequential Frequent Patterns is called as Content Patterns set and denoted by CP.

Stage 3: Each Tag associated with the blog may have multiple terms. After preprocessing, each tag is called as a Tag Pattern. The set of all Tag Patterns is denoted TP.

Stage 4: The set of Content Patterns (CP) and Tag Patterns (TP) can be combined in three ways to obtain the final Blog Patterns, denoted by BP. They are as follows:

Method 1: By applying union operation on CP and TP, i.e., BP = CP union TP

Method 2: By applying intersection operation on CP and TP, i.e., BP = CP intersection TP

Method 3: By applying the formula below: BP = (CP intersection TP) union ((CP-TP) in TP) union ((TP-CP) in Terms)

Both Methods 1 and 2 are straightforward. While Method 1 does a union of CP and TP to obtain BP (also called as Union Patterns), Method 2 does an intersection of CP and TP to obtain BP(also called as Intersection Patterns). Method 3 is quite involved. It is based on the idea that each content pattern or tag pattern may have one or more terms in it. Initially, the intersection of CP and TP is taken and added to the BP set. Now, the set (CP-TP) is considered. For each pattern of this set, each term of that pattern is checked to see if it is appearing in some Tag Pattern. If yes, then only that pattern is added to the BP set. The idea of checking (CP-TP) with all Tag Patterns is to verify whether the frequent patterns extracted are either relevant or not. Then, the set (TP-CP) is considered. For each pattern of this set, each term of that pattern is checked to see if it is appearing in the Terms set, which represents all the terms in the document. If yes, then only that pattern is added to the BP set. The idea of checking (TP-CP) with the Terms set is based on that fact that certain patterns may not be frequent but still represent the topic of the blog. Thus, Method 3 appears to be stricter than union (Method 1) and more lenient than intersection (Method 2).

In order to explain all the three methods, consider the following example:

Say, CP = $\{A, B, C, D\}$ and TP = $\{C, D, E, F\}$ where A, B, C, D, E and F are all patterns. Since each pattern may have more than one term, consider A = [a1, a2], B = [b1, b2, b3], C = [c1, c2], D = [d1, d2], E = [e1, e2], and F = [f1, b2, b3] and Terms = $\{a1, a2, b1, b2, b3, c1, c2, d1, d2\}$.

Method 1: When the union operation is applied, BP = $\{A, B, C, D, E, F\}$

Method 2: When the intersection operation is applied, BP = $\{C, D\}$

Method 3: When the third approach is taken,

- Patterns C and D are added to BP since they occur in both CP and TP (intersection).
- Pattern A in CP is not added to BP because none of A's terms (a1 and a2) are appearing in any Tag Pattern.
- Pattern B in CP is added to BP because majority of B's terms (b2 and b3) are appearing in Tag Pattern F.
- Pattern E in TP is not added to BP because none of E's terms are appearing in Terms Set.
- Pattern F in TP is added to BP because majority of F's terms (b2 and b3) are appearing in Terms set.
- So, the final BP set will be BP = $\{B, C, D, F\}$.

Blog Patterns obtained after combining Content Patterns and Tag Patterns can be considered as features representing the blog document. While union operation (Method 1) will increase the number of features and intersection operation (Method 2) will decrease the number of features, Method 3 tries to achieve a balance between the two. Blog Patterns obtained through Method 3 are expected to represent a blog in a better way when compared to Content Patterns, Tag Patterns, Union Patterns, and Intersection Patterns.

4 Evaluation of Blog Patterns

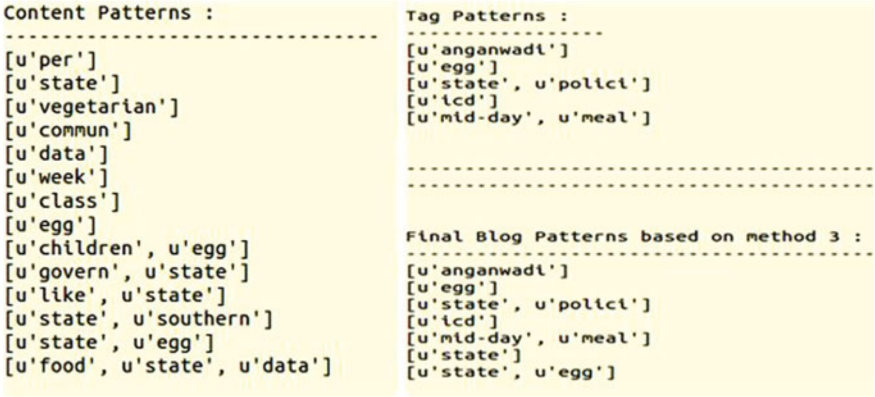
Now that the Content Patterns and Tag Patterns are combined to obtain the Blog Patterns, these Blog Patterns have to be evaluated to determine whether they represent the blog in a better way. Evaluation method used for this is clustering of documents and measuring the purity of clustering. That is, blog documents are clustered based on different patterns such that blogs that have common topics are grouped into the same cluster.

K-means clustering is used as a baseline clustering technique and purity of clusters obtained based on Content Patterns, Tag Patterns, Union Patterns, Intersection Patterns, and Blog Patterns—Method 3 can be compared. For k-means clustering, each blog document is represented as tf*idf vector (term frequency—inverse document frequency) and Euclidean distance measure is used. Individual terms are extracted from the patterns. The number of paragraphs in which a term occurs in a document gives the tf value of that term in that document. The relevance of that term for clustering is given by idf, which is based on terms occurrence in the document collection. The well-known formula is used for computing idf. Suppose, if there is a term that is a part of any pattern but it is not at all present in the document (because of tags), its term frequency is taken to be 1. In this way, tf*idf vectors are computed for each document and then clustering is applied.

5 Experimental Results

Blogs related to different topics can be collected from different Web sites. Each blog will have text content, tags, title, author, and much other information. The blog content and blog tags are then preprocessed by applying Stopwords Removal and Stemming techniques. Content Patterns, Tag Patterns, and Blog Patterns are then extracted based on the proposed methods. For example, consider a blog which is about the consumption of eggs in midday meals across different states in India. The Content Patterns, Tag Patterns, and the final Blog Patterns obtained based on Method 3 are given in Fig. 2. It can be seen that the proposed method removes redundant and noisy patterns from Content Patterns and Tag Patterns and retains those patterns that are meaningful.

For clustering purpose, 20 blogs each about 2 famous personalities are collected. Clustering is done and Purity of clustering is calculated. Another 20 blogs about third personality is collected and clustering is done on the entire collection. Next, another 20 blogs about fourth personality is collected and again clustering is done



```
Content Patterns :
------------------------------
[u'per']
[u'state']
[u'vegetarian']
[u'commun']
[u'data']
[u'week']
[u'class']
[u'egg']
[u'children', u'egg']
[u'govern', u'state']
[u'like', u'state']
[u'state', u'southern']
[u'state', u'egg']
[u'food', u'state', u'data']
```

```
Tag Patterns :
----------------
[u'anganwadi']
[u'egg']
[u'state', u'polici']
[u'icd']
[u'mid-day', u'meal']


----------------------------------------
----------------------------------------


Final Blog Patterns based on method 3 :
----------------------------------------
[u'anganwadi']
[u'egg']
[u'state', u'polici']
[u'icd']
[u'mid-day', u'meal']
[u'state']
[u'state', u'egg']
```

Fig. 2 Results for sample blog

Table 1 Results of blog clustering

Type of patterns	Purity (in %) for dataset with		
	2 clusters	3 clusters	4 clusters
Content Patterns	57.30	56.81	49.50
Tag Patterns	60.50	66.24	60.34
Union Patterns	52.50	53.35	46.56
Intersection Patterns	58.50	63.30	60.04
Blog Patterns (method 3)	65.50	65.52	62.02

on the entire collection. The results of clustering done based on all 5 Methods are given in Table 1.

From the Table 1, it can be seen that Blog Patterns (Method 3) represent any blog in a better way when compared to Content Patterns, Union Patterns, and Intersection Patterns. In some cases, the Tag Patterns maybe better than Blog Patterns (Method 3). However, since the Tag Patterns are derived from tags which is given by either bloggers or readers, it maybe inconsistent. So, it is ideal to re-verify those tags by using the proposed Method 3.

6 Conclusion and Future Work

The proposed technique for combining Blog Content Patterns and Blog Tag Patterns is effective in removing the redundant patterns present in them. Further, the effectiveness of this technique in retaining all the required patterns is analyzed based on Blog Clustering. The results obtained show that the Blog Patterns of proposed method are better than Content Patterns and Tag Patterns to represent a blog document. The proposed method can be applied to Blogs which have text content only. However, certain blogs may also contain other media (image, video, etc.) embedded into them by bloggers. The effect of these new dimensions can also be analyzed in determining the topics of blogs.

References

1. O'Leary DE (2011) Blog mining-review and extensions. Decision Support Systems, 51(4): 821–830.
2. M. Polanyi, The Tacit Dimension, Routledge & Kegan Paul, London, UK, 1966.
3. Yin Zhang, Kening Gao, Bin Zhang, Jinhua Guo, Feihang Gao and Pengwei Guo, Clustering Blog Posts Using Tags and Relations in the Blogosphere, 1st International Conference on Information Science and Engineering (ICISE), 2009.
4. F.S. Tsai, A tag-topic model for blog mining, -Expert Systems with Applications, 38 (5) (2011), pp. 5330–5335.
5. Y.-H. Chen, E.J.-L. Lu, M.F. Tsai, Finding keywords in blogs: Efficient keyword extraction in blog mining via user behaviors, Expert Systems with Applications, 41 (2) (2014), pp. 663–670.
6. Ning Zhong, Yuefeng Li and Sheng-Tang Wu, Effective Pattern Discovery for Text Mining, IEEE Transactions on Knowledge and Data Engineering, 24 (1) (2012).
7. Yuefeng Li, Abdulmohsen Algarni, Mubarak Albathan, Yan Shen, and Moch Arif Bijaksana, Relevance Feature Discovery for Text Mining, IEEE Transactions on Knowledge and Data Engineering, 27 (6) (2015).
8. Duc-Thuan Vo and Cheol-Young Ock, Learning to classify short text from scientific documents using topic models with various types of knowledge, Expert Systems with Applications, 42 (3) (2015), pp. 1684–1698.
9. Xiang Wang, Xiaoming Jin, Meng-En Chen, Kai Zhang, and Dou Shen, Topic Mining over Asynchronous Text Sequences, IEEE Transactions on Knowledge and Data Engineering, 24 (1), (2012).
10. Chenghua Lin, Yulan He, Richard Everson, Member, IEEE, and Stefan Ruger, Weakly Supervised Joint Sentiment-Topic Detection from Text, IEEE Transactions on Knowledge and Data Engineering, 24 (6) (2012).
11. Shenghua Bao, Shengliang Xu, Li Zhang, Rong Yan, Zhong Su, Dingyi Han, and Yong Yu, Mining Social Emotions from Affective Text, IEEE Transactions on Knowledge and Data Engineering, 24 (9) (2012).

Medicinal Side-Effect Analysis Using Twitter Feed

Priyanka S. Mane, Manasi S. Patwardhan and Ankur V. Divekar

Abstract As the use of social media network has been increasing, people tend to share health-related information on social sites. Twitter is used by large number of users and it is a wide source of information to analyze the drug related side effect. In this paper, we have developed an approach to analyze the contents of tweets to identify the adverse effects of a drug. An annotated dataset is used to train SVM classifier to identify the tweets showcasing medicinal side effects. The use of feature selection and dimensionality reduction techniques have allowed us to enhance the performance of the classifier in terms of accuracy by 10.34% as well as efficiency by nearly 66.31% as compared to the previous similar approaches.

Keywords Twitter · Adverse effects · Chi-square · Information gain · PCA

1 Introduction

Social media networks can serve as a source to analyze individual interests and their effects on personal life as they provide huge amount of data. Out of all the social networks, Twitter has a large number of users sharing different information. Twitter allows user to become a source of knowledge and expertise about certain topic. According to statistics, 500 million tweets are generated everyday and 200 billion tweets are generated every year. This statistics show that twitter contains large amount of data which can be used as a source of analysis.

P.S. Mane (✉) · M.S. Patwardhan
Department of Computer Science, Vishwakrma Institute of Technology,
Pune, India
e-mail: manepriyanka48@gmail.com

M.S. Patwardhan
e-mail: manasi.patwardhan@vit.edu

A.V. Divekar
Department of Electrical Science, Savitribai Phule Pune University, Pune, India
e-mail: ankur.divekar@cloudmoyo.com

P.K. Sa et al. (eds.), *Progress in Intelligent Computing Techniques: Theory, Practice, and Applications*, Advances in Intelligent Systems and Computing 719, DOI 10.1007/978-981-10-3376-6_7

There are traditional approaches that use national reporting tools engaging patients, practitioners, and researchers. These allow patients to voluntarily submit the reports to about the medicinal errors. The national reporting tool such as IHI Trigger tool [Institute for Healthcare Improvement] measures adverse drug events and allows conducting a review of patient records using trigger to identify possible side effects. In United States, the voluntary reporting system such as Patient Safety Network (PSNET) has been developed by U.S. Department of Health and Human services [1] that allows healthcare professionals to submit the cases that highlight medical errors and adverse effects. As the use of social networks is emerging, patients use various social networks such as Twitter, blog, and various forums to provide a review of medicines and other pharmaceutical information. As the twitter is used by millions of users, there are high possibilities that people are using twitter more often than any other reporting tool. Thus, we claim that social media such as Twitter is going to be a very good source of information for determining medicinal side effects.

The analysis of twitter data has the problem of high dimensionality. High dimensional data increases the complexity of data mining algorithms exponentially. The high dimensional data requires large volume of space. As the dimensions increase, the space requirement increases. Also, such high dimensional data is sparse meaning most of the feature-values (term frequencies) are zeros or not applicable for a given Tweet.

In this paper, our approach builds a classifier which takes twitter messages as input and separate outs twitter messages containing the adverse effect caused by a drug. We focus on the problem of high dimensionality in the twitter data. We use feature reduction techniques such as chi-square and information gain to improve the accuracy of the classifier, by selecting only contributing features (terms) and getting rid of negatively affecting features. The dimensionality reduction technique such as Principal Component Analysis is applied to identify the set of reduced number of dimensions which improves the efficiency of the classifier not hampering the accuracy. We compare our approach with a similar approach in [2] to find out that with the reduced number of dimensions have achieved better performance in terms of accuracy as well as efficiency.

2 Related Work

There are various studies performed on detecting the adverse reaction using social networking sites. The authors in [3] identified the adverse effects by analyzing the user comments on medicinal sites patient forum. This study is partitioned into four different methodologies: extracting medical entity using lexicons by utilizing metamap tool, extracting adverse drug reactions using transductive SVM classifier, classification of comments based on personal experience or hearsay, and finally analyzed the ADR (Adverse Drug Reaction) by using FDA (Food and Drug Administrator) reporting tool. They have used the medicinal site where data is not very sparse. The results of this study are 78.3% precision and 69.9% recall, for an f-measure of 73.9% which is moderate.

Yang et al. [4] have made use of association mining techniques to extract the adverse drug effects by analyzing the MedHelp health community. They have used lexicon-based ADR matching techniques. The study is performed on 10 drugs and five predetermined adverse effects. They have compared the side effect from the site with the predetermined drugs and have not detected the unknown adverse effects.

[5] describes an approach to find potential adverse events by analyzing the content of user comments on medicinal site. They collect the data from medicine Web site Dailystrength [6] use association mining algorithm to find out the patterns in the comment and based on this, they have generated rules by association and frequent pattern mining. This approach is highly dependent on the training data set and detects the adverse reaction patterns based on the matching patterns in the training data.

Bian et al. [7] describe an approach to find drug users and potential adverse events by analyzing the content of twitter messages utilizing Natural Language Processing (NLP) and to build Support Vector Machine (SVM) classifiers. Their study uses five drugs that were in clinical trials. They classify the positive and negative drug users and consider tweets of positive drug users to recognize if the tweets contains side effect. The accuracy of the classifier is 74% so the performance of the classifier is moderate due to the noise existed in the twitter message. There are no efforts in reducing the noise and errors such as spelling errors and abbreviations, etc.

Jian and Zheng [8] have used machine learning-based classifier to classify personal experience tweets, and use NLM's (National Library of Medicine) MetaMap software to recognize and extract word phrases that belong to drug effects. They classify the tweets based on the personal pronouns and sentiment analysis if tweet is identified as personal experience tweet. The corpus is not annotated by the domain expert and it does not classify if the tweet contains the adverse effect.

[2] develops a binary classifier to classify the tweets based on if the tweet contains side effect information or not. They have used 74 medicines and extracted the tweets consisting of those medicines. The dataset is divided into three subsets. The accuracy of the classifier is estimated using SVM classifier. There are no efforts to improve the efficiency of the classifier. Thus, the performance of the classifier is moderate.

3 Methods

In this section, we elaborate on the various steps required for classification of the tweets. The twitter classification process includes dataset, preprocessing, feature extraction, reduction, and classification. Figure 1 provides an overview of our methodology.

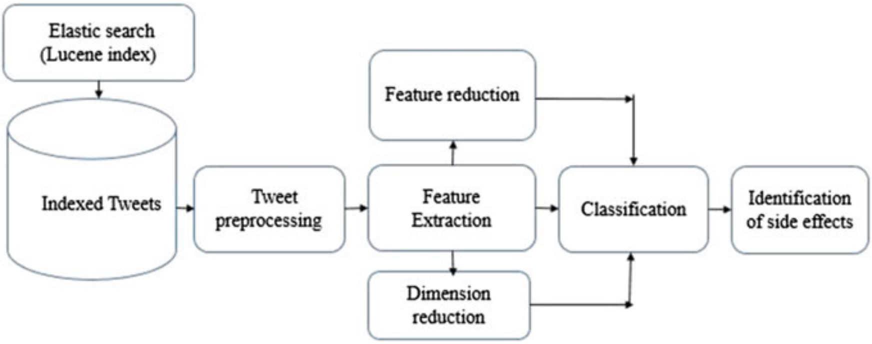


Fig. 1 Methodology

3.1 Dataset

We have used the annotated secondary twitter data provided by the authors Rachel and Pranoti [2]. This also helps us to provide a comparative analysis with this approach. The dataset includes 7000 tweets which can be used for data analysis. Out of these tweets, 1764 were removed from user timeline. We have used total 5236 tweets.

3.2 Elastic Search (Lucene Index)

The storage of dataset is important when there is a massive amount of data. The relational method of storage such as SQL performs slow search operations which is not significant in large amount of data. In our research, we have crawled the tweets and stored in elastic search which provides fast indexing and fast searching. Elastic search is a search server based on lucene [9]. Elastic search is also easily scalable, supporting clustering. It builds indices on dataset stored inside elastic search and retrieves the documents within a fraction of seconds.

3.3 Data Preprocessing

The twitter data is highly unstructured and informal; therefore, it contains more noise that is prone to errors. The twitter data includes slang words, abbreviations, and emoticons that need to be preprocessed to convert to meaningful form. We preprocessed the data to replace the slang words and abbreviation in the tweet, to remove the '@,' '#' and other special characters from the tweets, to remove the user name and re-tweets (RT).

After preprocessing, the text processing has been performed that includes removal of stop words, stemming, and tokenization. The stop words are removed by using stop words in NLTK toolkit in python [10]. All the words are stemmed to its root form and transformed to lower case and tokenized into individual words.

3.4 Feature Extraction and Reduction

For feature extraction, the bag of words model is used. After counting the occurrence of words in a tweet, the Term Document Frequency (TFIDF) of the terms is calculated that is used further for classification. Feature reduction techniques such as chi-square and information gain are applied to the dataset and most informative features are selected for classification.

3.5 Dimensionality Reduction

Dimensionality Reduction reduces the dimensions of the data and transforms into new dimensions. For dimension reduction, Principal Component Analysis (PCA) is used that orthogonally transforms the input dataset into new coordinates covering maximum variance of the given set of data points.

3.6 Classification

After feature extraction, reduction and dimension reduction, the classification is performed. The Support Vector Machine (SVM) is used for classification. It finds a hyperplane that separates the data into two different classes.

In our dataset, out of 5236 tweets there are 4376 tweets that do not contain side effects; whereas, 860 tweets contain the side-effects. Therefore, the dataset is highly imbalanced. To balance the dataset, it is partitioned to three subsets. The first dataset has equal no. of tweets from both the classes. The second dataset has 60% of tweets that do not contain side-effects and 40% of tweets that contain the adverse effects. The third dataset contains 70% of tweets without adverse effects and 30% of tweets with side-effects.

Classification is performed with all the features without any feature reduction and accuracy of classification is estimated which is compared with the classification accuracy gained after applying chi-square, information gain and PCA, taking only effective set of features/components into consideration.

4 Automatic Identification of Drug Side-Effect

To identify the side effects of a drug from tweets detected as having side effects, we use the ADR (Adverse Drug Reaction) based lexicon concepts. The ADR-based lexicon concepts are provided by authors in [2]. The lexicon concepts consists of concepts and UML (Unified Medical Language) concept IDs. The detected tweets are preprocessed to remove noise in the tweet. The stop words are removed, words are stemmed, and tokenization is performed. The tokens in the tweets are searched for matching tokens in the ADR lexical concepts. The string comparison with regular expression is performed to find the matching side-effect of a drug. The Table 1 shows few of the drugs and its extracted side-effects by using the procedure mention. Here, we have shown a few drugs and their side-effects for the demonstration purpose.

5 Results

In this section, we provide results in terms of accuracy and efficiency for four different scenarios. As our dataset is divided into three parts, the results are discussed here for second subset which consists of data 60% of data from one class and 40% of data from another class.

5.1 *Classification with All Set of Features*

After feature extraction, classification is performed on all set of features without any feature and dimensionality reduction. The tenfold stratified cross-validation is used to split the data into training and testing. The accuracy obtained with all 4117 features for second subset is 75.77% and the execution time is 283.065 s.

Table 1 Medicines and its extracted side-effects

Drugs	Side effects
Prozac	Insomnia, suicidal thoughts, feeling
Quetiapine	Abnormal dreams, feel like a
Seroquel	Drowsiness, restleness, weight gain
Trazodone	Headache, insomnia, hangover
Paxil	Anxiety, feel sick, weight gain
Effexor	Nausea, headache
Lamotrigine	Insomnia, feel sick, feel zombie

5.2 Classification After Applying Feature Reduction Using Information Gain and Chi-Square

The feature reduction technique chi-square test is performed on the features after feature extraction. The most informative features are selected using 5% level of significance. For second subset of data, 200 features are obtained as informative features using chi-square test. The classification accuracy obtained after chi-square test is 80.47% with 25.753 s execution time.

To select the most informative features at which the maximum accuracy is obtained information gain is used. Figure 2 demonstrates the graph of accuracy versus number of features.

The features are ranked based on the information gain. Thus in the above graph, on X axis the number of features depict those number of features having maximum information gain. As the number of features increase initially the accuracy increases, but it gets dropped in the later portion. The maximum accuracy is obtained at 780 features. Therefore, top 780 features are selected for classification after applying information gain.

In Table 2, we have compared the accuracy of our classifier after applying feature reduction with the accuracy in [2]. To measure the performance of the

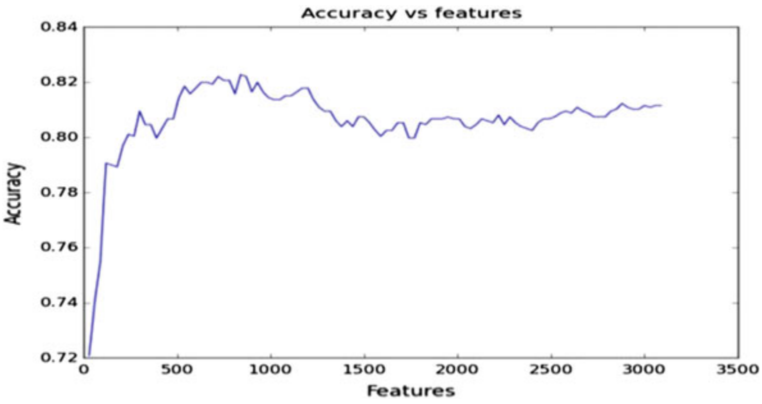


Fig. 2 Estimation of accuracy with number of features

Table 2 Classification results using SVM and feature selection algorithms

Accuracy of the classifier			
	First subset (%)	Second subset (%)	Third subset (%)
Accuracy in [2]	71.5	72.8	76.6
Chi-square	78.3	80.47	83.5
Information gain	82.47	83.61	85.61

classifier, 10 k stratified cross-validation is used. The accuracy in [2] is 72.8% for second subset. After applying feature reduction, the accuracy is increased to 80.47 and 83.61%.

5.3 Classification After Applying Dimensionality Reduction Using PCA

The accuracy remains constant for any number of components more than 1435 which is 75.78%. Therefore, only 1435 components are selected for classification. This results in performance gain in terms of efficiency. The execution time required before applying PCA is 283.065 s and after applying PCA is 127.70 s (Fig. 3).

5.4 Classification After Applying Feature Reduction and Dimensionality Reduction

In Fig. 4, the accuracy is varying from 80.87 to 80.75% up to 150 components. The accuracy is 81.11% that is maximum at component 170; therefore, 170 components can be selected for classification when PCA is applied after chi-square.

Figure 5 shows the number of features and accuracy when the PCA is applied after feature reduction using information gain. When PCA is applied, the accuracy is varying from 83.64 to 83.62% between 780 to 210 components and maximum is

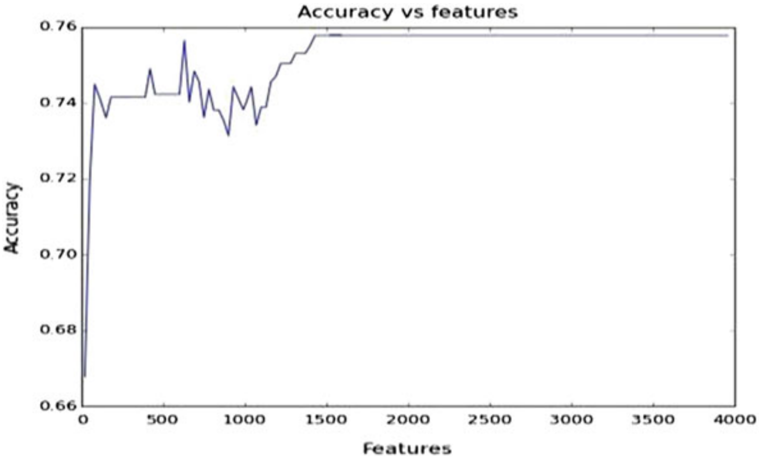


Fig. 3 Accuracy estimation without feature selection using PCA

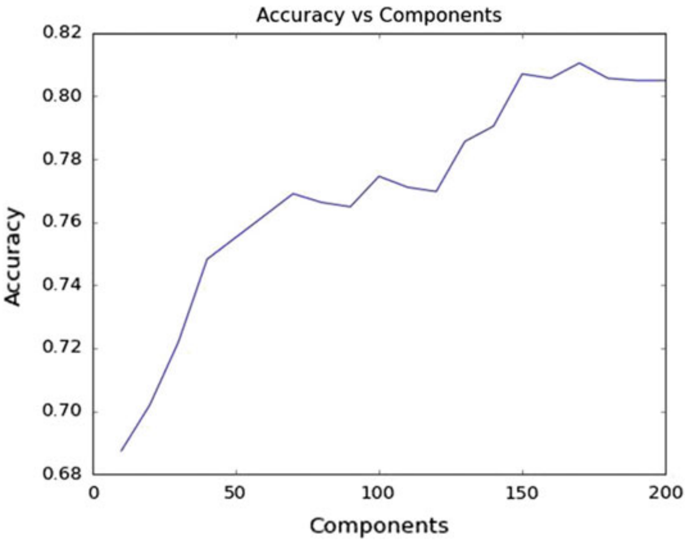


Fig. 4 Estimation of accuracy with different no. of components by applying PCA and chi-square

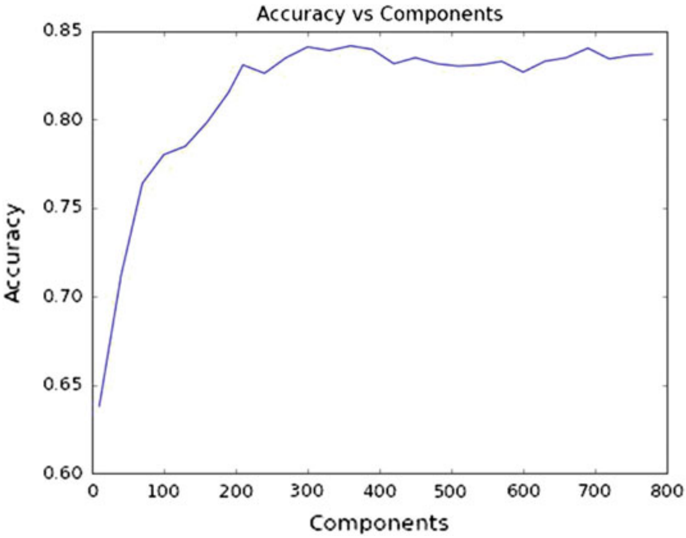


Fig. 5 Estimation of accuracy with different no. of components by applying PCA and information gain

Table 3 Summary of the classification

	Number of features or dimensions	Best accuracy (%)	Time required for execution (s)
Classification with all set of features	4117	75.77	283.065
Classification after applying feature reduction using Information Gain	780	83.61	123.82
Classification after applying feature reduction using Chi-square	200	80.47	25.753
Classification after applying dimensionality reduction using PCA	1435	75.78	127.70
Classification after applying Information gain and dimensionality reduction	300	84.21	41.713
Classification after applying Chi-square and dimensionality reduction	170	81.11	23.50

obtained at component 300 with 84.21% accuracy with an execution time of 41.713 s.

The Table 3 gives the summary of the methods applied in terms of the number of features or dimensions selected, best accuracy and execution time.

6 Conclusion

In this approach, we have developed a framework that classifies the tweets to identify adverse drug reactions. We use the SVM classifier to classify the tweets into tweets having information about side-effects. We have applied the feature reduction methods such as information gain and chi-square Test to improve the accuracy of the classifier from 72.8 to 80.47%. We reduce the dimension of the data using PCA from 200 to 170 after applying chi-square and from 780 to 300 after applying information gain, without hampering the classification accuracy. The efficiency of the classifier is improved by 60.31% due to application of PCA. Thus, for identifying tweets containing side-effects we have developed a technique with improved accuracy as well as efficiency.

References

1. Niraj L. Sehgal, MD, MPH; Sumant R. Ranji, MD; Kaveh G. Shojania, MD; Russ J. Cucina, MD, MS; Erin E. Hartman, MS; Lorri Zipperer, MA; Robert M. Wachter, MD: Development of a Web-Based Patient Safety Resource: AHRQ Patient Safety Network (PSNet).
2. Rachel Ginn, Pranoti Pimpalkhute, Azadeh Nikfarjam, Apurv Patki, Karen O'Connor, Abeed Sarker, Karen Smith and Graciela Gonzalez. Mining Twitter for Adverse Drug Reaction Mentions: A Corpus and Classification Benchmark. In proceedings of the Fourth Workshop

on Building and Evaluating Resources for Health and Biomedical Text Processing (BioTxtM2014). May, 2014. Reykjavik, Iceland.

3. Liu, X., and Chen, H. (2013). AZDrugMiner: an information extraction system for mining patient-reported adverse drug events in online patient forums. In Smart Health (pp. 134–150). Springer.
4. Yang CC, Jiang L, Yang H, Tang X. Detecting signals of adverse drug reactions from health consumer contributed content in social media. In: Proceedings of ACM SIGKDD workshop on health informatics. Beijing, China: ACM; 2012.
5. Nikfarjam, A., and Gonzalez, G. H. (2011). Pattern mining for extraction of mentions of adverse drug reactions from user comments. In *AMIA Annual Symposium Proceedings* (Vol. 2011, p. 1019). American Medical.
6. http://www.dailystrength.org/.
7. Bian J, Topaloglu U, Yu F. Towards large-scale twitter mining for drug-related adverse events. In: Proceedings of the 2012 international workshop on smart health and wellbeing. Maui, Hawaii: ACM; 2012. p. 25–32.
8. Jiang, K, and Zheng, Y. (2013) Minig Twitter Data for Potential Drug Effects. In Advance Data Mining and Application (pp. 434–443). Springer.
9. Sematext. Elasticsearch refresh interval vs indexingperformance. http://bit.ly/1iZoPGc, July 2013.
10. Edward Loper and Steven Bird NLTK: YThe Natural Language Toolkit Department of Computer and Information Science University of Pennysylvania, Philadelphia,19104–6389, USA.

A Study of Opinion Mining in Indian Languages

Diana Terezinha Miranda and Maruska Mascarenhas

Abstract Opinion mining is an area in natural language processing that is concerned with the determination of the opinion conveyed in a document text by a computer rather than through human intervention. This is extremely beneficial since it saves on resources required to inspect the document manually. Most of the research work done in this field is restricted to the English language. Opinion mining in Indian languages poses several challenges to researchers. This paper outlines some of the works done in this field for Indian languages and provides a brief description on each of them.

Keywords Opinion mining · Sentiment recognition · Indian languages · Natural language processing

1 Introduction

The area of sentiment analysis (also called opinion mining) pertains to the application of the fields of natural language processing and text analysis in the extraction, identification, and characterization of the sentiment contained in the given materials [1]. It is one of the most fascinating areas in artificial intelligence where a computer tries to behave like a human brain and analyzes a document text in order to determine its polarity without any human intervention.

There is a lot of work done in this field when it comes to the English language. However, when we consider Indian languages particularly regional ones, we find that there is not much work done in them. As of the 2001 Census by the

D.T. Miranda (✉) · M. Mascarenhas
Department of Computer Engineering, Goa College of Engineering,
Ponda, Goa, India
e-mail: diamir310@gmail.com

M. Mascarenhas
e-mail: maruskha@gec.ac.in

P.K. Sa et al. (eds.), *Progress in Intelligent Computing Techniques: Theory, Practice, and Applications*, Advances in Intelligent Systems and Computing 719, DOI 10.1007/978-981-10-3376-6_8

Government of India, there are about 122 major Indian languages and 1599 other languages which are spoken in India.

Despite their increasingly popular usage in recent times in blogs, social networking sites, and online forums there are several issues that prevent researchers from working in this field. Inadequate online processing support, unavailability of document databases, lack of a common script, insufficient funds for research, and excessive effort required to understand the complexities of the language often deter analysis of Indian languages for natural language processing.

This paper tries to describe several techniques used in opinion mining for Indian languages in detail. Some languages like Hindi, Bengali, etc., have more work done as compared to other regional languages like Marathi, Manipuri, etc.

2 Opinion Mining in Hindi

Joshi et al. performed opinion mining on 250 movie reviews written in Hindi, which were collected from various blogs using three methods. In the first method, RapidMiner 5.0 was used to train an SVM classifier on the in-language documents. Thereafter, the classifier was used to classify new documents. In the next method, a machine translator (Google translator) was first used to translate each of the documents to English. Then the classifier was modeled based on the standard movie review corpus and later used to classify the translated English documents. In the third approach, a Hindi SentiWordNet called H-SWN was created on the lines of the English SentiWordNet. It used various steps such as stemming, stop word removal and multiple sense disambiguation to create a classifier that will find out the polarity of the document using the scores obtained from the H-SWN. After experimentation, they found out that the first method of using an SVM classifier on an in-language corpus produced the best results with an accuracy of about 78% [2].

Mittal et al. extended the work done by Joshi et al. They used negations and conjunctions to make the H-SWN more effective. The negation word inverts the sentiment of the word preceding it. A window of a certain size was created and the polarity of all the words within this window was inverted. Additionally, care was taken to distinguish between forward and backward negations. Conjunctions are words that relate two parts of a sentence called discourse segments. Two types of conjunctions were taken into consideration. Conj_After gave more preference to the discourse segment following the conjunction while the discourse segment preceding the conjunction was dropped. Conj_Infer also gave more weight to the discourse segment after the conjunction as it assumed that the latter part of the sentence was inferred from the former part of the sentence. The proposed algorithm achieved an accuracy of about 80% [3].

Bakliwal et al. created a subjective lexicon for the Hindi language with a dependency only on WordNet and a small seed list of 45 words along with their polarities. They collected 900 Amazon reviews that had a size of less than or equal to 25 words. The reviews were then translated to Hindi using Google Translator.

Then they created a temporary seed list T of all the initial 45 words. A final seed list F was also created which is initially kept empty. For each word w in T, which is encountered in the text, the word is delisted from T. If the word is in F, it is not populated any further but its polarity is added to the total polarity p of the text. But if w is not in F then it is added to F while all the synonyms and antonyms are added to T with the same and opposite polarities of w, respectively. These steps are performed iteratively till T becomes empty. This algorithm produced results with an accuracy of about 68% [4].

3 Opinion Mining in Bengali

Das and Bandyopadhyay developed SentiWordNet for the Bengali language using a English to Bengali dictionary and the original SentiWordNet in English, which is an automatically created lexical tool that allocates positive and negative scores to each WordNet synset [5]. This method returned 35805 Bengali entries. A supervised classifier was generated with the help of the lexicon and other features like positional aspects, etc. This classifier achieved a precision of 74.6% and a recall of 80.4%. Over the years, they incorporated several improvements to their original algorithm. An interactive game was created that returned the annotated words along with their polarities. A bilingual English to Bengali dictionary as well as synonym–antonym relations were exploited to find the polarities of words. Lastly, machine learning from a pre-annotated corpus was performed to ascertain the polarity of the given documents [6].

Several lexical clues like punctuation marks, emoticons, negations, discourses, reduplications, emoticons, and structural clues like rhetoric statements were used to obtain information regarding the emotions being conveyed by the documents. Thereafter, two methods were used for classification of the documents. The first method used a Conditional Random Field while a Support Vector Machine Classifier was used for the second method. It was found that the SVM classifier significantly outperformed the CRF classifier. However, both suffered from sequence labeling, and the label bias problem w.r.t. other non-emotional parts of a sentence. In another experiment, the CRF classifier was used to assign Ekman's six emotion class tags which are anger, fear, disgust, sad, happy, and surprise along with three types of intensities which are high, general and low to the Bengali blog documents. And the CRF classifier was able to successfully do this [7].

4 Opinion Mining in Tamil

Sudhakar and Bensraj used ANTLR to tokenize the document text into its constituent words and then each of the words was tagged using the POS tagger. In the next step, they performed word sense disambiguation and stemming. The keywords or words associated with emotions were spotted and mapped onto a tridimensional space which is also called as the circumflex. The circumflex defined the activation (stimulation of activity), valence (a positive or negative evaluation), and control (submissiveness) features that the keyword conveys. Then the average emotional dimensions for each input text was calculated based on the circumflex. Later the most appropriate sentiment label was selected based on the features that were extracted from the terms found in the text. This label was usually taken for a bag of words. Then using this classification, the document text was read out as speech with the emotions being conveyed in the text. For this text to speech conversion, phonetic analysis, prosodic modeling and intonation were used so that the text matched the emotion being conveyed through speech. For classification of the text fuzzy neural networks were used. Here the input data was converted into fuzzy data using triangular membership functions. The above experiment was performed using MATLAB and the FNN had produced significantly successful results [8].

Giruba et al. used a data set that contained documents classified under five domains namely politics, cinema, business, health, and sports. The Tamil morphological analyzer was used to retrieve nouns and verbs. The constituent term domain frequencies was calculated and inserted into a hash table. The keys of the hash table were the constituent terms and the term frequencies were the values in the hash table. Then the document was analyzed to determine if positive words occur in close proximity of negative words and vice versa. Accordingly, a score was assigned based on the negation elements. Next, a score was assigned based on the amount of pleasantness in the words. This was determined by the phonetic classification in Tamil along with the place and manner of articulation done. Four taggers were then used to tag the words namely noun tagger, verb tagger, Urichol tagger and case tagger. After applying several other steps, a supervised backpropogation network was constructed. However, since emotion recognition is to do with emotional intelligence, unsupervised learning was preferred and Hebbian learning was incorporated. For stories, the precision of the Neural Network was about 60% especially because of the subplots, which tend to give it a neutral sentiment. However, for lyrics and songs the precision was about 70% [9].

5 Opinion Mining in Manipuri

Nongmeikapam et al. used the conditional random field (CRF) to perform opinion mining in Manipuri. The data set consisted of letters to the editor published in few daily newspapers. POS tagging using CRF was done to identify the verbs and the lexicon of the verbs was modified to contain the polarity of the word. The polarity categories namely positive, negative, and neutral were assigned manually to each of the verbs. Based on this, the polarity of each sentence was calculated. The total count for each polarity category was summed up separately and the polarity with the highest count decided the polarity of the document text [10].

6 Opinion Mining in Telugu

Ekman’s six basic emotion types namely anger, sadness, happiness, disgust, fear, and surprise were used to tag the emotional words in a document set consisting of Telugu blog texts and English news data. The non-emotional words were tagged using a neutral type. Furthermore, language experts verified these tags. Additionally, emotion tags were assigned to sentences depending on the highest emotion score assigned to the constituent words of the sentences. Conditional random field (CRF) was used to do the classification of emotion and non-emotion words, thereby subsequently classifying the sentences as well. The CRF classifier used 10 active features to perform the classification task. The features included a Part of Speech (POS) tagger (classified words as noun, verb, adjective and adverb), first sentence in a topic, SentiWordNet emotion word (a word that existed in the Telugu or English SentiWordNet was assumed to contain emotion), reduplicated words, question words, special punctuation symbols, quoted words, emoticons, colloquial or foreign words and sentences with a maximum of eight words. The above evaluation was conducted on the Telugu and English data sets and the results were found to be satisfactory in both scenarios [11].

7 Opinion Mining in Kannada

Anil Kumar et al. created an exhaustive positive and negative keyword list. Then they applied a baseline algorithm to the document text, which used two counters—a positive counter and a negative counter initializing both counters to zero. The positive counter was incremented each time a word in the positive keyword list was encountered while the negative counter was decremented each time a word in the negative keyword list appeared in the document text. Later a list of negator words was compiled. The positive and negative counters were decremented and incremented, respectively each time a positive or negative keyword appears within a window containing a negator word. Then a POS tagging algorithm was used which computed the polarity of all the words which were tagged as adjectives. Additionally, Turney’s algorithm was used to improve the accuracy, which checked for certain POS tagging patterns for phrases containing two or three words. Any phrase in the document text that followed any of the given patterns was further analyzed and its polarity was computed. The number of keywords in each sentence along with its polarity was computed. The document text was assigned the polarity of the sentence, which contained the most number of keywords. If a tie occurred then the average polarity was considered as the polarity of the document [12].

Deepmala et al. also developed an opinion mining system for the Kannada language. Using a rule-based stemmer based on the Paice method, each word from the document was extracted and stemmed. A lexicon word list was created which contained the polarities for 5043 Kannada words where the polarities ranged from

−5 (very negative) to +5 (very positive). Each extracted word was compared with the polarity lexicon word list. The suffix for each common word was checked with the entries in the list of negation suffixes. If a match occurred, then the polarity of the word containing the negator suffix was negated. The polarity of each sentence was calculated as the sum of the polarities of the individual words in that sentence. The opinion conveyed by the document was considered positive if there were maximum positive sentences and negative if there were maximum negative sentences [13].

8 Conclusion

In conclusion, the study of sentiment analysis was done for several Indian languages. The most commonly used method was the conditional random field (CRF) classifier especially in cases where features that could be selected for the classifier were easy to identify.

Some languages such as Bengali and Hindi had comparatively extensive work done in this area. On the contrary, languages such as Telugu and Manipuri had less work done in the field of opinion mining.

Languages that already had sufficiently accurate processing tools such as POS taggers, stemmers, WordNet, etc., required much less investment in terms of time and money as compared to those languages, which had primitive language processing tools.

References

1. Kaur, A., Gupta, V.: A Survey on Sentiment Analysis and Opinion Mining Techniques. In: Journal Of Emerging Technologies In Web Intelligence, Vol. 5, No. 4, November, (2013).
2. Joshi, A., R., B. A., Bhattacharyya, P.: A Fall-back Strategy for Sentiment Analysis in Hindi: a Case Study. In: Proceedings of International Conference on Natural language Processing (ICON), Karagpur (2010).
3. Mittal, N., Agarwal, B., Chouhan, G., Bania, N., Pareek, P.: Sentiment Analysis of Hindi Review based on Negation and Discourse Relation. In: Proceedings of International Joint Conference on Natural Language Processing 2013, 45–50, Nagoya, Japan (2013).
4. Bakliwal, A., Arora, P., Varma, V.: Hindi Subjective Lexicon: A Lexical Resource for Hindi Polarity Classification. In: Proceedings of the Eight International Conference on Language Resources and Evaluation (2012).
5. Das, A., Bandyopadhyay, S.: Phrase-level Polarity Identification for Bangla. In IJCLA Vol. 1, No. 1–2, pp. 169–182, (2010).
6. Das, A., Bandyopadhyay, S.: SentiWordNet for Indian Languages. In: Asian Federation for Natural Language Processing, China, pp. 56–63, (2010).
7. Das, D., Bandyopadhyay, S.: Labeling Emotion in Bengali Blog Corpus – A Fine GrainedTagging at Sentence Level. In: Proceedings of the 8th Workshop on Asian Language Resources, August 21–22, Beijing, China, (2010).

8. Sudhakar, B., Bensraj, R.: An Efficient Sentence-based Sentiment Analysis for Expressive Text-to-speech usingFuzzy Neural Network. In: Research Journal of Applied Sciences, Engineering and Technology 8(3): 378–386 (2014).
9. Giruba Beulah, S.E. and Karky, M.: On Emotion Detection from Tamil Text.
10. Nongmeikapam, K., Khangembam, D., Hemkumar, W., Khuraijam, S., Bandyopadhyay, S.: Verb Based Manipuri Sentiment Analysis. In: International Journal on Natural Language Computing (IJNLC) Vol. 3, No.3, June, (2014).
11. Manchala1, S., Chandra Mohan, D., Nagesh, A.: Word and Sentence Level Emotion Analyzation in Telugu Blog and News. In: International Journal of Computer Science, Engineering and Applications (IJCSEA) Vol.2, No.3, June, (2012).
12. Anil Kumar, K. M., Rajasimha, N., Reddy, M., Rajanarayana, A., Nadgir K.: Analysis of Users' Sentiments from KannadaWeb Documents. In: Proceedings of Eleventh International Multi-Conference on Information Processing (2015).
13. Deepamala, N., Ramakanth Kumar, P.: Polarity Detection of Kannada documents. In: Proceedings of IEEE International Advance Computing Conference (IACC 2015), BMSCE Bangalore, June 12–13 (2015).

Part II
Applications of Informatics

A Pragmatics-Oriented High Utility Mining for Itemsets of Size Two for Boosting Business Yields

Gaurav Gahlot and Nagamma Patil

Abstract Retail market has paced with an enormous rate, sprawling its effect over the nations. The B2C companies have been putting lucrative offers and schemes to fetch the customers' attractions in the awe of upbringing the business profits, but with the mindless notion of the same. Knowledge discovery in the field of data mining can be well harnessed to achieve the profit benefits. This article proposes the novel way for determining the items to be given on sale, with the logical clubs, thus extending the Apriori algorithm. The dissertation proposes the high-utility mining for itemsets of size two (HUM-IS2) Algorithm using the transactional logs of the superstores. The pruning strategies have been introduced to remove unnecessary formations of the clubs. The essence of the algorithm has been proved by experimenting with various datasets.

Keywords High-utility itemsets · Retail · Business yields · Log mining · Offers · Knowledge discovery · B2C companies · Pragmatics · Apriori algorithm

1 Introduction

The business organizations have evolved and also manifested the market approaches. For making a smooth progress of the business profits, the essential factors have been pointed out. Business intelligence is one among them. For a business firm, the decision making and its strategies for marketing play the vital roles. Business intelligence is a consolidation of warehousing, mining, querying, etc. All the decision support techniques have to be refined, along with the information extraction methodologies [1].

G. Gahlot (✉) · N. Patil
National Institute of Technology Karnataka, Surathkal, Mangalore 575 025,
Karnataka, India
e-mail: gauravgahlot299@gmail.com

N. Patil
e-mail: nagammapatil@nitk.ac.in

P.K. Sa et al. (eds.), *Progress in Intelligent Computing Techniques: Theory, Practice, and Applications*, Advances in Intelligent Systems
and Computing 719, DOI 10.1007/978-981-10-3376-6_9

Retailing is give and take relationship between the individuals, or organizations. The retail attempts to fulfill the customers' needs and wants. It is the brands and the companies making those goods or products to sell for developing their profits. In the end, it comes out to be the revolution about those companies and brands [2]. All the leading market companies like Walmart, Amazon, Walmart, Big Bazaar, etc., have taken the revolution to different apex, under the crust of making money benefits.

The behaviors of the customers of the retail stores is key centric absorbent for imparting the services. Due to growing competition among the companies, the customers fall in dilemma to choose from. There psychology makes the companies lure the customers [3]. And here arises the concept of putting schemes or offers on the items by the companies.

Transactional logs of the supermarkets give the transactions history made by the customers. The customer's behavior and the transaction logs have an overlapping characteristics. Decision support systems facilitate the realistic view of the data which would else have been overlooked. By proper analysis of these logs using the Log mining, the company officials can take crucial decisions [4]. These can be regarding what all items to be kept on schemes and offers for putting in clubs by the companies.

Apriori algorithm [5] lies at the roots of the discovery of the itemsets. The main motive is to find frequent patterns within the dataset. It is based upon the *Support-Confidence Framework*. *Support* acts as the frequency threshold for each item or itemsets.

The paper proposes a modification to Apriori algorithm called high utility mining for itemsets of size two (HUM-IS2). It works upon the utility parameter to find the clubs of two item products, which can be sold on schemes, using the pruning strategies for improving the performance.

2 Related Work

Data mining has outstretched its span, sprawling the significance in areas like web, e-learning, shopping, etc. Han et al. [6] describe the mining mechanisms in various data formats and scenarios, such as graphs, multimedia, text, web, transactional databases, etc. In [4], Zhang et al. propose the methodology for prediction of the behaviors of the customers based on the transaction rules within. Their proposed Mafia algorithm is an extension to find maximum frequent itemsets with the common user itemsets.

Cai et al. [7] gave an idea for weighted association rule concept, addressing the shortcoming of support-confidence framework. Here, the weights of the items act as the importance of them. The issue is regarding the negligence of the quantity of items in the transaction. Yan et al. [8] researched for the web recommendation system. They also considered the number of times the particular page was visited to find the dominating web pages.

The evolutionary meta heuristic algorithms have also been worked upon by many researchers. ARMGA [9], QuantMiner [10], GENAR [11], and G3PARM [12] are some of the researches. These algorithms use the concept of chromosomes to find the relationship between the antecedent and consequent. Luna et al. [12] have utilized the algorithm for context-free grammar for extracting the quantitative association rules.

Wang et al. [13] gave the concept of ontological interference in the mining process. This marked the upcoming of logical perspections instead of the statistical approaches.

All the previous researches done use the frequency parameter. But, the occurring frequency for an itemset may not demonstrate the truth to be a enough parameter for the significance. This is because the frequency reflects to the transactions pertaining to an itemset. This does not help in relieving the *utility* of the same. The utility can be any of the cost or price, the profit, the revenue, or some other preferred expression.

Also, it might happen that frequent itemsets contribute less to the entire benefit, while infrequent itemsets contribute more. Any business is more attracted for identifying the most valuable users, i.e., the customers of high profit fraction contribution. The following example illustrates that support-confidence framework may mislead the decision makers.

Example 1 Example 1 Considering the transactional logbase D as depicted in Table 1(a). It has nine transactions from T_1 to T_9 and eight items from i_1 to i_8. The sales quantity of each item for a particular transaction has been exhibited in the round brackets. Table 1(b) gives the profits for every product in unit sale. For the item set $\{i_4, i_5, i_6\}$, the support and utility when calculated using the Table 1(a), (b) comes out to be 4 and 36, respectively. Hence, the profit of the item set is 36. The total utility of the individual items, i_4 and i_6 is 22 and 20 respectively. When observed for the contribution of each of these two items for the item set, it marks 8 for i_4 and 16 for i_6. Using the statistical approach will mislead the business managers to take wrong decision regarding which product drives the other one. The actual relationship is associated by i_6 to $\{i_4, i_5\}$. This means that the selling of the entire item set $\{i_4, i_5, i_6\}$ contributes a big portion to the utility of i_6 to 16 on 20. Therefore, the frequency is not considered to be a sufficient indicator for knowing whether a particular itemset is of huge impact or not.

Table 1 Example

(a)

T_{id}/$Trans_{id}$	Transaction
$Trans_1$	$i_1(4), i_3(1), i_5(6), i_6(2)$
$Trans_2$	$i_4(1), i_5(4), i_6(5)$
$Trans_3$	$i_2(4), i_4(1), i_5(5), i_6(1)$
$Trans_4$	$i_4(1), i_5(2), i_6(6)$
$Trans_5$	$i_1(3), i_3(1), i_5(1)$
$Trans_6$	$i_2(1), i_6(2), i_8(1)$
$Trans_7$	$i_4(1), i_5(1), i_6(4), i_7(1), i_8(1)$
$Trans_8$	$i_4(7), i_5(3)$
$Trans_9$	$i_7(10)$

(b)

Item	Utility
i_1	3
i_2	4
i_3	5
i_4	2
i_5	1
i_6	1
i_7	2
i_8	1

3 Proposed System

Utility is the prime attribute for defining the "usefulness" in terms of the profits. The utility mining goals at identifying the high utility itemsets which drive the majority portion in the total utility. The utility-confidence model has its applicability in various applications. To obtain the profit, the manager can take decision for rewarding the customers purchasing more than some threshold and give a lucrative deduction on the bill amount or shipping cost. Therefore, considering the necessity to have some pragmatics based mining mechanism, *Utility Mining* came up. The key terms in the formal definitions for utility mining are as follows:

- $I = \{i_1, i_2, ..., i_m\}$ is the *itemset.*
- $D = \{T_1, T_2, ..., T_n\}$ is the *transactional log* entry. $T_i \epsilon D$.
- $o(i_p, T_q)$ quantifies i_p for T_q. For instance, $o(i_1, T_1)$ is 4.
- $s(i_p)$ represents cost of i_p. As an example, $s(i_1)$ is 3.
- $u(i_p, T_q)$ being the *utility* is product outcome of $o(i_p, T_q)$ and $s(i_p)$. Considering Table 1, $u(i_1, T_1)$ is $4 \times 3 = 12$.
- $u(X, T_q)$ is the utility for X in T_q. It is $\sum_{i_p \epsilon X} u(i_p, T_q)$.
- *u(X)*, utility for the itemset X is $\sum_{T_q \epsilon D \wedge X \subseteq T_q} u(X, T_q)$.
- $tu(T_q)$, utility for T_q is $u(X, T_q)$. Here, X denotes itemset of all items in T_q.

3.1 HUM-IS2 Algorithm

If the utility of an itemset X meets with the minimum threshold specified, then it is a *High Utility Itemset*. The Apriori property for pruning the search region for candidate itemset cannot be put up directly for mining due to neither non-monotonicity nor monotonicity of the utility constraint. For reduction of the search space and enhancement of the performance, the *Transaction-Weighted Utility* (TWU) concept is used satisfying the downward closure property.

TWU of itemset X is $\sum_{T_q \epsilon D \wedge X \subseteq T_q} tu(T_q)$. For X, if TWU(X) is not meeting the utility threshold, then all its supersets will also be of low utility.

Assume there is a known total sequence ordering α amongst all the items in the database. Consequently, if one item i occurs before other item j in that sequence, the relation is denoted by $i \; \alpha \; j$. For $\forall j \; \epsilon \; X$, if $i \; \alpha \; j$, we say $i \; \alpha \; X$; X being an itemset. This ordering is applied for enumerating the itemsets without duplication. Now onwards, for the paper an itemset will be considered in an order. Particularly, it is a series of items in increasing manner of TWU values. Lexicographic order is applied for items bearing same TWU values.

It becomes a necessity for a procedure to mine HUIs of size-2. For avoiding the computation overhead and multiple scans of the database, Liu et al. [14] have shown a vertical mining approach for generating high utility item sets in a single phase. It deploys the data structure named as *Utility List* for an itemset. It runs a DFS in a preorder manner from left direction toward the right, and the utility list of k-itemset is formed by that of *(k-1)*-itemsets.

For an itemset X, the set of items following X in T_q is represented by T_q/X. The utility list of X is shown as UL(X), shown in Table 2(a). For X, the UL comprises the tuples (T_q, *iutil*, *rutil*) for every T_q having X. *iutil* is the sum of the utilities of all items in T_q/X. *rutil* is $\sum_{i \epsilon T_q/X} u(i,T_q)$. For a itemset, if sum of *iutils* < *min_util*, then it is of low utility. For *min_util*, if the addition of the *iutil's* and *rutil's* in utility list is less than it, there is no extension Y of X to be referred as a HUI. This is called promising utility of X.

3.2 Co-occurence Based Pruning Strategies

For reduction of joins for evaluation of UL for itemset P_{xy}, the *Estimated Utility Co-occurrence Pruning* (EUCP) strategy is used for eliminating a low utility extension of P_{xy} and its extensions. It encompasses the TWU pruning strategy to prune itemset of size greater than two. The TWU value of all itemsets of size 2 is saved

Table 2 Example Continued: (a) Utility Lists and (b) EUCS of promising items (c) PUCS of i_1, i_2

(a)

Item	Utility List
i_7	(7,2,7),(9,20,0)
i_2	(3,16,8),(6,4,2)
i_1	(1,12,13),(5,9,6)
i_3	(1,5,8),(5,5,1)
i_4	(2,2,9),(3,2,6),(4,2,8),(7,2,5),(8,14,3)
i_6	(1,2,6),(2,5,4),(3,1,5),(4,6,2),(6,2,0),(7,4,1)
i_5	(1,6,0),(2,4,0),(3,5,0),(4,2,0),(5,1,0),(7,1,0),(8,3,0)

(b)

Item	i_7	i_2	i_1	i_3	i_4	i_6
i_2	0					
i_1	0	0				
i_3	0	0	40			
i_4	9	24	0	0		
i_6	9	30	25	25	54	
i_5	9	24	40	40	71	79

(c)

PUCS of i_1			PUCS of i_2		
Item	i_3	i_6	Item	i_4	i_6
i_6	25		i_6	24	
i_5	40	25	i_5	24	24

in 2D-array structure called, *Estimated Utility Co-occurrence Structure* (EUCS). It being a store of rows in the manner (a, b, c) $\epsilon I' \times I' \times R$ such that $\text{TWU}(a, b) = c$. For itemset *X* and item *i*, if there exists no row (X_k, y, c) complying c < *min_util*, then *Xy* and its supersets are low utility itemsets.

The utility value given to the EUCS values by *i* lacks impression after its subproblem is completed. This provides a *Promising Utility Co-occurrence Structure* (PUCS) for every item for reduction of candidate itemsets. The PUCS of an item *a*, shown by $PUCS_a$, is a collection of rows in the manner (b, c, v), $a\ \alpha\ b$ and $a\ \alpha\ c$, $D_{ab} \neq \phi$; and $D_{ac} \neq \phi$; such that *v* being sum of *e.iutil* and *e.rutil* where *e* is item in UL(*a*). Tables 2(b), (c) give the EUCS and PUCS structures for Table 1.

The proposed HUM-IS2 algorithm (Algorithm 1) follows the sequential approach. Using the utility list, HUIs will be found. Then using pruning concepts, the itemsets will be made minimal. Algorithm HUM-IS2 builds the necessary data structures and parameters for carrying out the processing. It also initiates the finding of the clubs of items. Procedure *FindU* supports the Algorithm HUM-IS2 by developing the Utility List of the itemsets. Procedure *SearchExtensions* checks the other extra areas, i.e., the itemset clubs which can be searched here itself for calling as HUI or not. Procedure *Club* is used to validate the club formed using the decisions of EUCS and PUCS. All procedures help Algorithm HUM-IS2 to carry the processing.

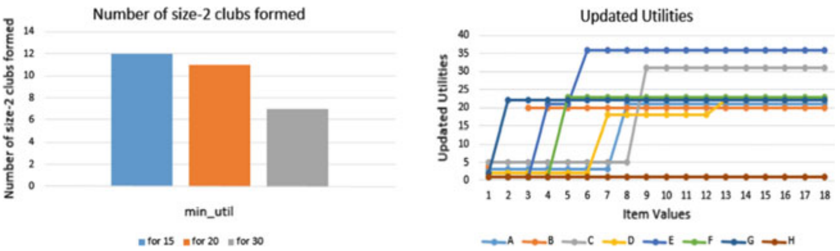
Algorithm 1: HUM-IS2 Algorithm

```
Data: D: Transactional Dataset, min_util:Minimum Utility
Result: High Utility Itemsets
C:{}
for every item i in D do
    find the $TWU_i$;
    if $TWU_i \geq min\_util$ then
        C=C ∪ i;
Put C in increasing manner of TWUs;
Delete i from D bearing TWU≱ min_util;
Construct UL of every sustaining item i;
Build EUCS structure;
Build PUCS structure;
for every item i in C do
    if $\sum UL_i.iutils \geq min\_util$ then
        i becomes the highUtilityItemset;
        Save i with its utility value;
    if $\sum UL_i.iutils + \sum UL_i.rutils \geq min\_util$ then
        tail:{}
        for every item j in C, i α j from left to right do
            if ∃ (i,j,v) ϵ EUCS with $v \geq min\_util$ then
                Compute X={i} ∪ {j};
                Compute $UL_X$ = FindU($UL_i$, $UL_j$);
                Compute tail=tail ∪ $UL_X$;
        Call SearchExtensions(i,tail)
return High Utility Itemsets
```

```
Procedure FindU
Data: $UL_x$:Utility List of item x,$UL_y$:Utility List of item y
Result: $UL_{xy}$:Utility List of item {x,y}
$UL_{xy}$=ϕ,EU=0 for every element $e_x$ ϵ $UL_x$ do
    if ∃ $e_y$ ϵ $UL_y$ and $e_x$.tid is equal to $e_y$.tid then
        Compute $e_{xy}$ as ( $e_x$.tid,$e_x$.iutil+$e_y$.iutil,$e_y$.rutil;
        Compute EU=EU+$e_x$.iutil+$e_y$.iutil;
        $UL_{xy}$=$UL_{xy}$ ∪ $e_{xy}$;
return $UL_{xy}$
```

```
Procedure SearchExtensions
Data: x: item,tail: variable for extensions
Result: High Utility Itemsets
if tail=ϕ then
    return;
for every P ϵ tail from right to left do
    if $\sum UL_P$.iutils ≥ min_util then
        if P.utility<$\sum UL_P$.iutils then
            Save P with its utility value as $\sum UL_P$.iutils;
    if $\sum UL_P$.iutils + $\sum UL_P$.rutils ≥ min_util then
        tailNew:{}
        for every item S ϵ tail, P α S from left to right do
            if Club(x,P,S)==True then
                Compute Z={P} ∪ {S};
                Compute $UL_Z$ = Find_U($UL_P$, $UL_S$);
                Compute tailNew=tailNew ∪ $UL_Z$;
        Call SearchExtensions(x,tailNew)
```

```
Procedure Club
Data: x,P,S:all being the items
if $EU_S$ then
    if size of P = 2 then
        if ∃ (P,S,v) ϵ EUCS with v < min_util then
            return False;
    else
        if ∃ (P,S,v) ϵ $PUCS_x$ with v < min_util then
            return False;
return False
```



(a) Number of size-2 itemsets formed having high utility.

(b) The flow of the items for attaining their utility gradually.

Fig. 1 Results

4 Experimental Evaluation

The dataset mentioned in Table 1 has been used. The real dataset used is the Retail dataset collected by frequent itemset mining dataset repository [15]. On applying the algorithm on the synthetic dataset, following graphs have been plotted.

Figure 1a describes the number of size-2 itemsets formed having high utility. As the iterations proceed, the item combinations are processed. The itemsets formed are checked for their utility value. In this figure, different minimum utility thresholds were taken and the graph has been plotted against the number of size-2 itemsets found. Figure 1b depicts the way the items attain their utility gradually while executing the program. As and when the algorithm flows, the utility of the items keeps on changing. It may happen that the utility for an item may take many steps as is in the case of the item ‘D’ while may remain stagnant like for ‘H’. This happens because of the buying nature of the customers.

For Retail dataset, minimum utility was kept as 75 and then, the num of size-2 clubs formed were 15.

5 Conclusion

Association rule mining is carried upon rigorously by scientists and researchers to find the associations between itemsets using concepts of statistics like support and confidence. But, there happens to be no semantic implication, which can be directed by utility. For this integration, the utility has been used to mine itemsets attaining high utility. The algorithm, on comparison to Apriori, works faster and also gives semantic relationship within the items on the basis of utility.

References

1. Gang, Tong, Cui Kai, and Song Bei. “The research & application of Business Intelligence system in retail industry.” Automation and Logistics, 2008. ICAL 2008. IEEE International Conference on. IEEE, 2008.
2. Anand, Akshay, and Snigdha Kulshreshtha. “The B2C adoption in retail firms in India.” Systems, 2007. ICONS’07. Second International Conference on. IEEE, 2007.
3. El-Deen Ahmeda, Rana Alaa, et al. “Performance Study of Classification Algorithms for Consumer Online Shopping Attitudes and Behavior Using Data Mining.” Communication Systems and Network Technologies (CSNT), 2015 Fifth International Conference on. IEEE, 2015.
4. Zhang, Yanyu, and Yonggong Ren. “A Method of Predicting Users’ Behaviors Based on Intertransaction Association Rules.” Web Information Systems and Applications Conference, 2009. WISA 2009. Sixth. IEEE, 2009.
5. Agrawal, Rakesh, and Ramakrishnan Srikant. “Fast algorithms for mining association rules.” Proc. 20th int. conf. very large data bases, VLDB. Vol. 1215. 1994.
6. Han, Jiawei, Micheline Kamber, and Jian Pei. Data mining: concepts and techniques. Elsevier, 2011.
7. Cai, Chun Hing, et al. “Mining association rules with weighted items.” Database Engineering and Applications Symposium, 1998. Proceedings. IDEAS’98. International. IEEE, 1998.
8. Yan, Liang, and Chunping Li. “Incorporating pageview weight into an association-rule-based web recommendation system.” AI 2006: Advances in Artificial Intelligence. Springer Berlin Heidelberg, 2006. 577–586.
9. Yan, Xiaowei, Chengqi Zhang, and Shichao Zhang. “ARMGA: identifying interesting association rules with genetic algorithms.” Applied Artificial Intelligence 19.7 (2005): 677–689.
10. Salleb-Aouissi, Ansaf, Christel Vrain, and Cyril Nortet. “QuantMiner: A Genetic Algorithm for Mining Quantitative Association Rules.” IJCAI. Vol. 7. 2007.
11. Mata, J., J. L. Alvarez, and J. C. Riquelme. “Mining numeric association rules via evolutionary algorithm.” ICANNGA’01, Proceedings of the 5th international conference on artificial neural networks and genetic algorithms, Prague, Czech Republic. 2001.
12. Luna, José M., José Raúl Romero, and Sebastián Ventura. “Design and behavior study of a grammar-guided genetic programming algorithm for mining association rules.” Knowledge and Information Systems 32.1 (2012): 53–76.
13. Xuping, Wang, Ni Zijian, and Cao Haiyan. “Research on association rules mining based-on ontology in e-commerce.” Wireless Communications, Networking and Mobile Computing, 2007. WiCom 2007. International Conference on. IEEE, 2007.
14. Liu, Mengchi, and Junfeng Qu. “Mining high utility itemsets without candidate generation.” Proceedings of the 21st ACM international conference on Information and knowledge management. ACM, 2012.
15. Frequent Itemset Mining Implementations Repository, http://www.fimi.cs.helsinki.fi/data/.

ILC-PIV Design for Improved Trajectory Tracking of Magnetic Levitation System

Vinodh Kumar Elumalai, Joshua Sunder David Reddipogu, Santosh Kumar Vaddi and Gowtham Pasumarthy

Abstract This paper puts forward the hybrid control algorithm, which integrates the iterative learning control (ILC) scheme with proportional integral velocity (PIV) control, for improved trajectory tracking of magnetic levitation system. ILC is a type of model-free controller, which is used for systems that perform repetitive tasks. Adjusting the control inputs based on the error information obtained during previous iterations, ILC tries to enhance the transient response of the closed-loop system. One of the striking features of ILC is that even without the full dynamic model of the plant, it can yield perfect trajectory tracking by learning the plant dynamics through iterations. Adopting this learning control feature of ILC, this paper aims to synthesize ILC with PIV for both improved tracking and better robustness compared to conventional PIV. The efficacy of the proposed ILC-PIV controller framework is assessed through a simulation study on the magnetic levitation plant for reference following application.

Keywords ILC · PIV · Intelligent control · Magnetic levitation system · Command following

V.K. Elumalai (✉) · J.S. David Reddipogu · S.K. Vaddi · G. Pasumarthy
School of Electrical Engineering, Vellore Institute of Technology, Vellore 632014, India
e-mail: vinodhkumar.e@vit.ac.in

J.S. David Reddipogu
e-mail: joshua.reddipogu@vit.ac.in

S.K. Vaddi
e-mail: vaddi8143@gmail.com

G. Pasumarthy
e-mail: gowthampasumarthy1996@gmail.com

P.K. Sa et al. (eds.), *Progress in Intelligent Computing Techniques: Theory, Practice, and Applications*, Advances in Intelligent Systems and Computing 719, DOI 10.1007/978-981-10-3376-6_10

1 Introduction

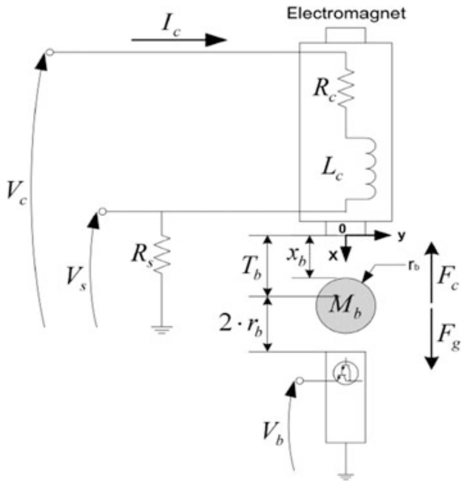
ILC, which adapts the concept of learn from mistakes, is a type of learning control strategy that is used for systems which perform a given task repetitively. By learning the system dynamics through iteration, ILC aims to improve not only the trajectory tracking performance of the system but also the robustness [1]. Unlike classical feedback and feedforward control methods, ILC utilizes the knowledge of the control signal from previous iterations to perform perfect command following of periodic signals and to reject the periodic disturbances. The key feature of ILC is that it is a model-free and adaptive control strategy. Hence, even without knowledge of the full dynamics of the system, ILC can yield satisfactory tracking performance by exploiting the error information obtained during previous iterations [2]. Introduced by Arimoto in 1984, ILC has been utilized in many of the engineering applications which are repetitive or cyclic in nature. For instance, Maeda et al. [3], combined ILC with disturbance observer for rejecting repeated disturbances in automated excavation. Chen and Hwang [4], assessing the performance of ILC on a pneumatic actuated X-Y table for position tracking application, reported that ILC can yield satisfactory performance at different speeds. Motivated by the robustness of ILC, Kim et al. [5], through numerical simulation, validated the performance of ILC on spatially interconnected systems, whose complete system dynamics is uncertain. Similarly, Chen et al. [6], using continuous sliding mode technique, put forward a robust ILC and assessed the performance on a benchmark rotary servo plant. Xu et al. [7] compared and analyzed the two of the fundamental ILC versions namely previous cycle learning and current cycle learning. Readers can refer to [8] for overview and survey of different results reported on ILC.

Taking advantage of the model-free and adaptive nature of ILC, we aim to synthesize the current cycle feedback (CCF) ILC with the PIV controller, for improved command tracking and robustness of the system against exogenous disturbance. The conventional PIV controller is a model based controller and requires accurate plant model for perfect control. However, the plant model may contain uncertainty due to various reasons including modeling error and actuator saturation. Moreover, the periodic disturbances present in the system may degrade the performance of the PIV control. Hence, to enhance the robustness of the PIV scheme against disturbance and model uncertainty, we put forward an ILC-PIV controller framework. Three test cases namely, nominal tracking, tracking under exogenous disturbance, and tracking during model uncertainty are assessed through a simulation study.

2 Magnetic Levitation System Description

The magnetic levitation (maglev) system comprises an electromagnet, a steel ball, and a photo sensory circuit. The entire system is encased in a rectangular enclosure and the control objective is to make the ball levitate and follow the input command.

Fig. 1 Schematic diagram of magnetic levitation system



The ball position can be manipulated by controlling the current given to the coil. Figure 1 shows the schematic diagram of maglev system. For brevity and page constraints, instead of presenting the complete modeling of the maglev, the following numerical transfer function model and the system parameters of the magnetic levitation plant are borrowed from [9].

$$G(s) = \frac{x_b(s)}{I_c(s)} = -\frac{K_b \omega_b^2}{s^2 - \omega_b^2} \tag{1}$$

As one of the open loop poles of the system is placed on the right half of the s-plane, feedback control is necessary. Hence, in the following section PIV is designed for stabilizing the plant and ILC is integrated for improving the tracking and robustness of the closed-loop system.

3 ILC-PIV Control Scheme

Figure 2 shows the block diagram of CCF-ILC integrated with PIV control scheme. The conventional PIV control scheme is to stabilize the plant and ILC is to improve the tracking and robustness of the closed-loop scheme. Unlike the conventional PID control in which the velocity of the error is given as a feedback signal, PIV control feeds back the measured velocity, which can result in reduced overshoot in the closed-loop response. The key advantage of PIV control over PID is that it can eliminate the derivative kick created by the abrupt change in set point. Thereby, PIV focuses more on smooth set point tracking rather than disturbance rejection. Hence to supplement the control for better disturbance rejection ILC is integrated using CCF technique, which is a type of ILC update strategy that updates the

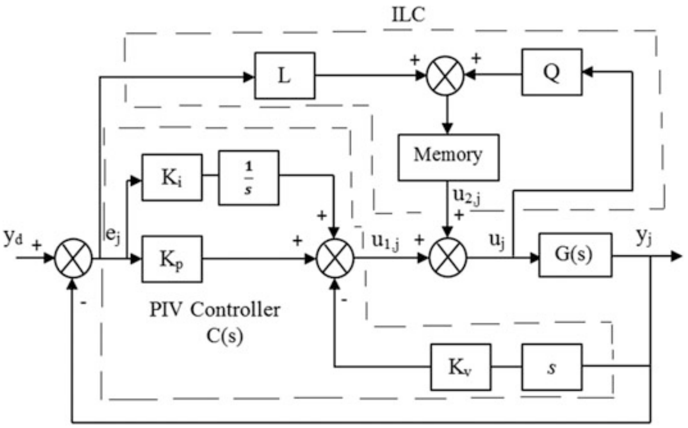


Fig. 2 ILC-PIV control scheme

control law based on the error at the current iteration. The following section gives the key expressions of the ILC-PIV synthesis for trajectory following application.

The PIV control law is given by

$$u_{1,j} = k_p(y_d - y_j) + k_i \int (y_d - y_j)\, dt - k_v \frac{dy_j}{dt} \tag{2}$$

In PIV, instead of the derivative of the error signal, the direct derivative of the output is taken to minimize the sudden deviation in output due to abrupt change in input.

3.1 CCF-ILC

CCF-ILC uses the error information available from the present control trial to update the subsequent control input such that the tracking error in the next iteration is minimized. Hence, current cycle tracking error (e_{j+1}) is included in the updating law of CCF-ILC.

$$u_{j+1} = Qu_j + Le_j + Ce_{j+1} \tag{3}$$

where Q and L indicate the ILC filters, and C represents the feedback controller. The current cycle tracking error is given by

$$e_{j+1} = y_d - y_{j+1} \tag{4}$$

Therefore, the control update becomes,

$$u_{j+1} = Qu_j + Le_j + Cy_d - Cy_{j+1} \tag{5}$$

$$u_{j+1} = Qu_j + Le_j + Cy_d - CGu_{j+1} \tag{6}$$

$$u_{j+1} = (1 + CG)^{-1}(Q - CG)u_j + y_d(1 + CG)^{-1}(L + C) \tag{7}$$

The learning rule will converge if it satisfies,

$$\left|u_{j+2} - u_{j+1}\right| \le \delta\left|u_{j+1} - u_j\right| \; with\, \delta < 1 \tag{8}$$

The convergence rate will increase if we select,

$$(1 + CG)^{-1}(Q - CG) \ll 1 \tag{9}$$

Theoretically, for the ILC to converge after the first trial,

$$(Q - CG) = 0 \tag{10}$$

Therefore, the L filter is given by:

$$L = QG^{-1} \tag{11}$$

The input output relation between y_d and y is

$$\frac{y}{y_d} = \frac{(1 + Q)^{-1}(L + C)G}{1 + (1 + Q)^{-1}(L + C)G} \tag{12}$$

Substituting (11) into (12) results in

$$\frac{y}{y_d} = \frac{Q + CG}{1 + CG} \tag{13}$$

The Q filter is introduced mainly to compensate for the deviation in the plant model. For asymptotic convergence of tracking error, the Q filter should have “low pass” characteristics. Hence,

$$Q(s) = \begin{cases} 1 & \omega \in (0, \omega_c) \\ 0 & \omega > \omega_c \end{cases} \tag{14}$$

4 Results and Discussion

The performance of the ILC-PIV control scheme is tested using Matlab/Simulink. The gains of the PIV controller are chosen based on the pole placement technique such that the response to a step input yields an overshoot of less than 10% and a settling time of 1 s. The respective gains of the PIV are $K_p = -227.23$, $K_v = -3.78$ and $K_i = -192.32$. One can note that the gains are negative due to the positive feedback loop in the maglev model. For the CCF-ILC, the Q filter is designed as a low pass filter with a cut off frequency of 57.18 rad/sec, which is same as the natural frequency of the plant. Hence, the second order low pass Q filter is given by,

$$Q(s) = \frac{3270}{s^2 + 114.36s + 3270}$$

The L filter determined based on the inversion technique given in (11) is,

$$L(s) = \frac{-3270s^2 + 1069000}{22.83s^2 + 2611s + 74650}$$

4.1 Nominal Tracking

To assess the reference following performance of the ILC-PIV control, a sinusoidal test signal with a peak to peak amplitude of 2 mm at a frequency of 0.5 Hz is given as a reference signal. From Fig. 3, which shows the tracking response and error of the control scheme, it can be noted that the controller yields a maximum tracking error of 0.009 during transient state and guarantees asymptotic convergence.

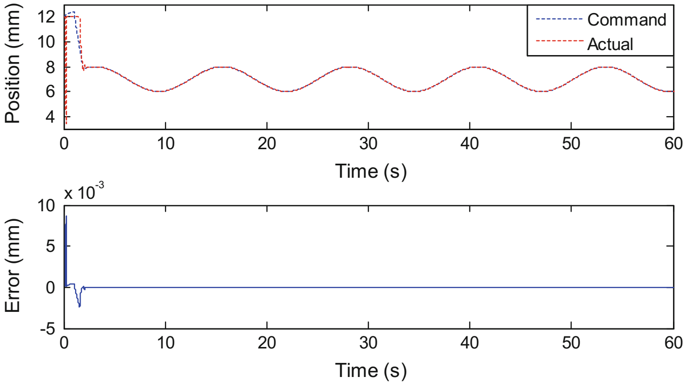


Fig. 3 Nominal set point tracking

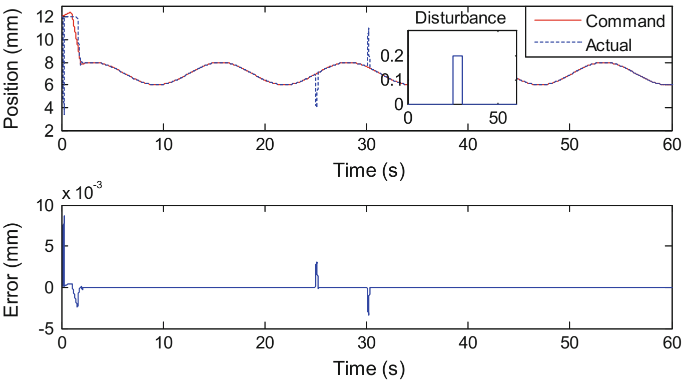


Fig. 4 Command tracking during exogenous disturbance

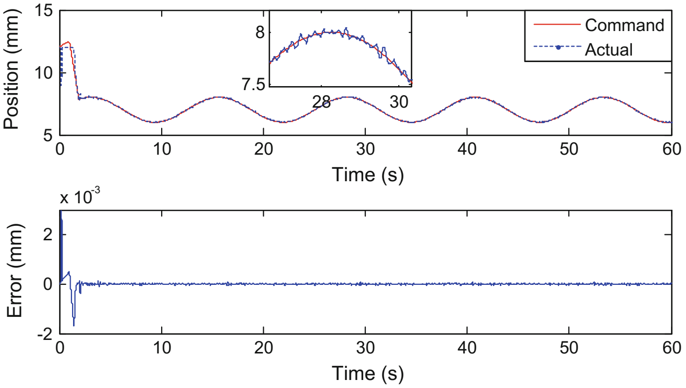


Fig. 5 Tracking during model uncertainty

4.2 Disturbance Rejection

A pulse disturbance with an amplitude of 0.2 mm is introduced from t = 25 s to t = 30 s to assess the regulatory response of the control scheme. It can be noted from Fig. 4, which shows the response of the system during exogenous disturbance, that the deviation in state trajectory is quickly brought to the desired trajectory within 2 s. Also, the peak value of the tracking error during the disturbance is kept within 0.003 mm.

4.3 Model Uncertainty

The model of the plant is assumed to be varying and a test case of 10% uncertainty is introduced at the plant level. Figure 5 shows the tracking response of the

ILC-PIV hybrid scheme during model uncertainty and a zoomed in view of the trajectory is also included to highlight the deviation in trajectory. It is worth noting that the control scheme can yield satisfactory tracking with minimum level of oscillation.

5 Conclusions

In this paper, the ILC-PIV hybrid control strategy for improved command tracking and robustness of the magnetic levitation system is presented. The motivation behind synthesizing ILC with PIV is the need for enhancing the robustness of the conventional PIV control scheme against disturbance. Using CCF method, the control law of ILC is updated and the Q and L filters are designed based on the inversion technique. Through numerical simulations, the set point tracking performance of ILC-PIV controller framework is validated. In addition, the ability of the control scheme to tolerate the exogenous disturbance and model uncertainty are also assessed. Simulation results substantiate that integrating the ILC with PIV can significantly improve the robustness and reference following capability of the closed-loop system.

References

1. Wu, Y., Zou, Q., Su, C.: A Current Cycle Feedback Iterative Learning Control Approach for AFM Imaging. IEEE Trans. Nanotechnol. 8 (4) (2009) 515–527.
2. Eglence, M.: Iterative learning control vs. feedback control - an experimental study. Report Nr. 025CE2002, Faculty of EEMCS, University of Twente, (2002).
3. Maeda, G.J., Manchester, I.R., Rye, D.C.: “Combined ILC and Disturbance Observer for the Rejection of Near-Repetitive Disturbances With Application to Excavation”, IEEE Trans. Control Syst. Technol. 23 (5) (2015) 1754–1769.
4. Chen, C.K., Hwang, J.: Iterative learning control for position tracking of a pneumatic actuated X–Y table. Control Eng Pract 12 (2005) 1455–1461.
5. Kim, B.Y, Lee, T., Kim, Y.S., Ahn, H.S.: Iterative learning control for spatially interconnected systems. Appl Math Comput 237 (2014) 438–445.
6. Chen, W., Chen, Y.Q., Yeh C.P.: Robust iterative learning control via continuous sliding-mode technique with validation on an SRV02 rotary plant. Mechatronics 22 (2012) 588–593.
7. Xu, J.X., Lee, T.H., Zhang, H.W.: Analysis and comparison of iterative learning control schemes. Eng Appl of Artif Intel 17 (2004) 675–686.
8. Wang, Y., Gao, F., Doyle, F.J.: Survey on iterative learning control, repetitive control, and run-to-run control. J Process Contr 19 (2009) 1589–1600.
9. Kumar, E.V, Jerome, J.: LQR based optimal tuning of PID controller for trajectory tracking of Magnetic Levitation System. Procedia Eng 64 (2013) 254–264.

Performance Analysis and Optimization of Spark Streaming Applications Through Effective Control Parameters Tuning

Bakshi Rohit Prasad and Sonali Agarwal

Abstract High-speed data stream processing is in demand. Performance analysis and optimization of streaming applications are hot research areas. Apache Spark is one of the most extensively used frameworks for in-memory data stream computing and capable of handling high-speed data streams. In streaming applications, controlling, and processing of data streams for optimized and stable performance within the available resources is of utmost requirement. There are various parameters that can be tuned to achieve the optimum performance of streaming applications deployed on Spark. This work explores the performance of stream applications in the light of various tunable parameters in Spark. Further, a relationship among the performance response and controlling parameters is established using linear regression. This regression model enables the prediction of performance response before actual deployment of a streaming application. The work determines an interrelationship between block interval and number of threads for optimized performance of streaming application also.

Keywords Performance optimization • Apache spark • Data stream • Streaming application

1 Introduction

In present scenario of computing technologies, the stream computing covers a wide range of applicability in various domains. Applications such as real-time monitoring in health care domain, intrusion tracking system, web-click stream mining, high-speed stream log analysis, outlier or anomaly detection, etc., process continuous arriving data streams for various purposes like machine learning, patter

B.R. Prasad (✉) · S. Agarwal
Indian Institute of Information Technology, Allahabad, India
e-mail: rohit.cs12@gmail.com

S. Agarwal
e-mail: sonali@iiita.ac.in

P.K. Sa et al. (eds.), *Progress in Intelligent Computing Techniques: Theory, Practice, and Applications*, Advances in Intelligent Systems and Computing 719, DOI 10.1007/978-981-10-3376-6_11

identification, analytics, recommendation, etc. Such applications need high-speed data stream processing frameworks. Aurora [1] massive online analysis (MOA) [2], S4 [3], storm [4], scalable massive online analysis (SAMOA) [5, 6], Apache Spark [7], etc., are recently developed frameworks that are well suited for developing stream processing applications. Apache Spark is a cluster computing framework that supports the applications having iterative jobs using working sets [7, 8]. Apache Spark uses resilient distributed data (RDDs), which are distributed collection of data and key abstraction of the framework [9]. High-speed processing power of Spark is achieved by its in-memory computation strategy, which makes it suitable for real-time stream processing [10]. A Spark streaming application needs to be tuned for best performance.

As a broader goal, performance considers two important aspects; average scheduling delay and average processing time. Thus, the objective of performance tuning is to achieve:

- Reduced scheduling delay through efficient exploitation of cluster resources and reduced processing time which is required to process the batch of data.
- Selection of appropriate batch size and level of parallelism so that the processing of data batches can cope up with the receiving rate to keep the system stable.

Various aspects are possible for achieving optimized performance [11]. In this research work, the focus is on tuning the data processing parallelism and block intervals for specific batch interval. This research work aims to achieve following three research outcomes:

- Performance analysis of streaming application in light of control parameters such as block interval, number of threads, and batch interval.
- Establishing a regression model capable of predicting the performance of streaming application before it is actually deployed.
- Exploring interrelationship between block interval and number of threads for optimized performance.

The paper is divided into five sections. The proposed methodology for performance analysis of streaming applications is described in Sect. 2. Further, it also elaborates Spark control parameters based on proposed model. In Sect. 3, the obtained results are shown along with suitable plots. Related research work is discussed in Sect. 4. Concluding remarks with future scope of the work is presented in Sect. 5.

2 Proposed Methodology

To achieve the target research outcomes mentioned in previous section, a proposed system model is shown in Fig. 1.

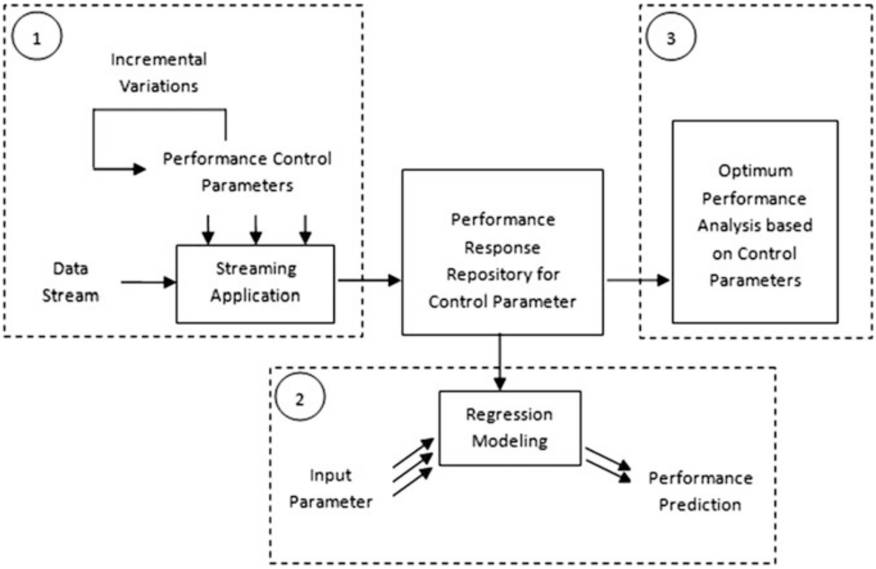

1
Incremental Variations
Performance Control Parameters
Data Stream
Streaming Application
Performance Response Repository for Control Parameter
3
Optimum Performance Analysis based on Control Parameters
2
Regression Modeling
Input Parameter
Performance Prediction

Fig. 1 System model for performance analysis

The system model has three modules each of which performs necessary processing to obtain the desired objectives. Module 1 is responsible for obtaining the first research objective. It runs a streaming application with varying configurations of performance control parameters, i.e., batch interval, block interval, and number of threads. The performance of application is observed in terms of average scheduling delay and average processing time. These observations are stored in a common repository for further analysis and processing. Module 2 achieves the second goal of this work via fitting a linear regression model on the stored observations. It applies ordinary least square linear regression modeling considering performance control parameters as independent variables (i.e., predictor variables) whereas average scheduling delay and average processing time as dependent variable (i.e., response variable). Module 3 performs analytical calculations on the stored observations in order to establish an interrelationship between numbers of threads and block interval for different configurations of batch intervals. Thus, it finds the configuration of number of threads with respect to block interval that gives the optimized performance of the streaming application thereby accomplishes the target of this work.

2.1 Spark Performance Tuning Parameters

There are various parameters of an application that are configurable in Spark. The performance of the application in Spark depends on the settings of these parameters. Some of these parameters are common to various applications supported in Spark. On the other hand, some of the parameters are specific to the streaming applications such as batch interval, block interval, etc. [12]. These Spark parameters can be tuned to control behavior of Spark application thereby the overall performance of application. Some of the parameters used for experiments performed in this work are listed in Table 1 along with their significance and default values in various settings.

2.2 Experiment Setup

In this work, we used a well-known and well-established Spark streaming application 'NetworkWordCount', that reads streaming data through a socket stream connection. As part of experimental settings, system and Spark configurations are listed in Table 2. Dataset is taken from web access logs for HTTP requests received by NASA [13].

Table 1 Spark tuning parameters

Spark tuning parameters	Description of parameter
Spark.streaming.batchInterval	Sets the interval of the batch of data stream to be read
Spark.streaming.blockInterval	Sets interval at which the read data stream is chunked into various data blocks and then stored in Spark
Spark.streaming.receiver.maxRate	It defines the maximum rate (records/second) at which any stream can receive data
Spark.streaming.unpersist	It causes forced removal of RDDs that are persisted in Spark's memory
Spark.default.parallelism	It specifies the default number of partitions done in RDDs
Spark.storage.memoryFraction	Specifies the fraction of Java Heap that is allocated for the purpose of memory caching of Spark
Spark.executor.memory	Sets the memory amount to be used per executor process
Spark.driver.memory	It configures the amount of memory to be allocated to driver process
Spark.executor.cores	If there are several cores then this setting facilitates many executors on each worker node for an application

Table 2 System setting for streaming application

System/spark configuration	Parameter value
#Machines = 2	Intel(R) Core(TM)-i3 CPU
Processor Speed	3.30 GHz
Total number of cores	8
RAM	6 GB
Operating System	64-bit Linux (Ubuntu 12.04)
Spark Version	1.2.0
Spark.executor.memory	6 GB
Spark.executor.cores	8 (as default value, all cores are used)
Spark.streaming.batchInterval	Varying {1 s, 2 s,….}
Spark.streaming.blockInterval	Varying {200, 250, 300, 350, 400,450, 500, 550}
local[n]	n is set as {2, 4, 6, 8,…..}

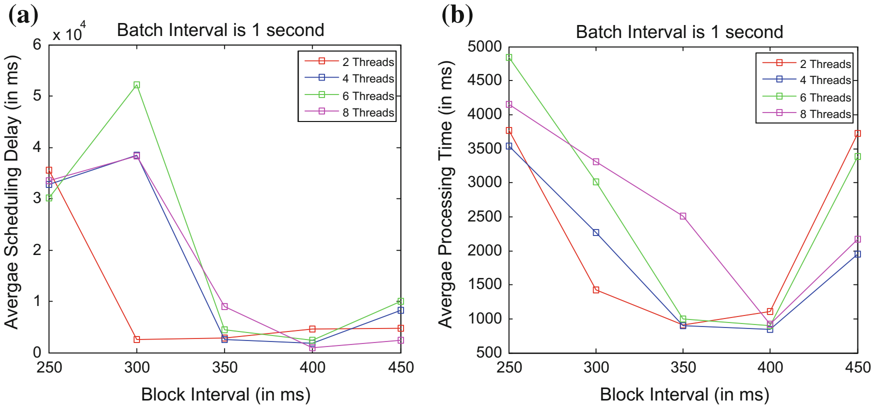


Fig. 2 Performance assessment plots with respect to batch interval 1 s, **a** For average scheduling delay, **b** For average processing time

3 Results and Analysis

The graphs shown in Fig. 2a, b is plotted for variations in average scheduling delay and average processing time, respectively, with respect to the changes in batch interval for 1 s. Different line series in each of the graphs represent the relationship for varying number of threads, i.e., 2, 4, 6 and 8. It is evident that when we increment the block interval, average scheduling delay and average processing time decreases; and decrements up to a certain point. Then after, they begin to rise.

Similar experiments are repeated with batch interval 2 s and same kind of behavior is found. It establishes the fact that the performance of any streaming application is best achieved at a certain block interval for specific setting of number of threads used. Optimum value of average scheduling delay and processing time for different number of threads for batch interval 1 and 2 s are listed in Table 3.

Table 3 Optimum average scheduling delay and average processing time

Batch interval (sec)	Number of threads	Average scheduling delay (in ms)		Average processing time (in ms)	
		Optimum value	At block interval	Optimum value	At block interval
1	2	2801	350	906	350
	4	1753	400	847	400
	6	2403	400	896	400
	8	934	400	922	400
2	2	1598	450	1268	450
	4	2815	500	1956	500
	6	2770	500	1343	500
	8	1817	500	1241	500

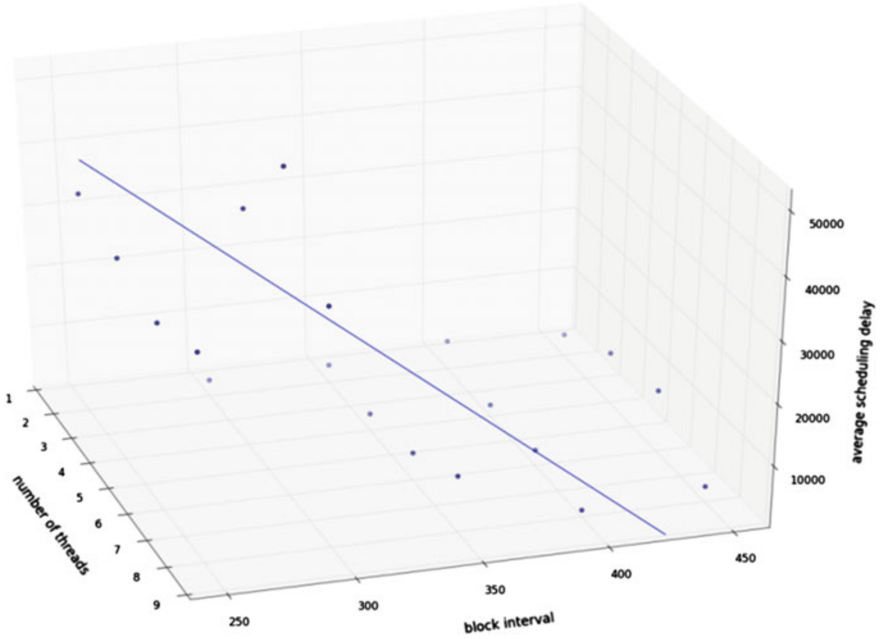


Fig. 3 Regression plot of average scheduling delay for batch interval 1 s

To predict scheduling delays and processing time before actually deploying the application in industries, enterprises, etc., is of utmost requirement. To obtain relationship between performance and controlling parameters, we used regression line fitting technique [14, 15]. Figure 3 describes the best regression line among the controlling parameters, i.e., number of threads, block interval, and average scheduling delay for batch interval of 1 s. Similarly, Fig. 4 depicts the best regression line among the controlling parameters and average processing time.

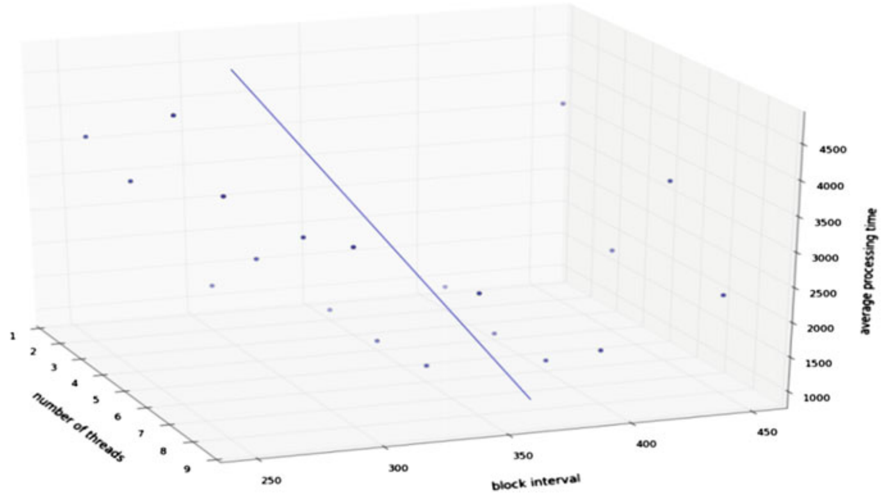


Fig. 4 Regression plot of average processing time for batch interval 1 s

Table 4 Regression coefficient analysis

Regression predictor terms	ASD			APT		
	Coefficient	t	P-value	Coefficient	t	P-value
RC_{BAI}	1.3808	0.262	0.045	1.4633	2.858	0.007
RC_{NT}	−145.2887	−4.128	0.000	−13.9032	−4.073	0.000
RCB_{LI}	352.6950	0.317	0.033	278.5650	2.580	0.014
C	6.351e + 04	4.442	0.000	4372.6000	3.153	0.003

The regression line is fitted using ordinary least square method with varying block intervals, number of threads and batch intervals taken as independent parameter and average scheduling delay and average processing time taken as dependent parameter. The obtained regression coefficients and other parameters for plotted regression line are listed in Tables 4 and 5 respectively.

RC_{BAI}, RC_{NT}, RC_{BLI}, and C represent regression coefficients corresponding to batch interval, number of threads, block interval, and constant, respectively. ASD and APT represent average scheduling delay and average processing time, respectively. Based on these parameter and coefficients, the obtained regression line equation can be written as given in Eqs. (1) and (2).

$$ASD = 1.3808 \times RC_{BAI} - 145.2887 \times RC_{NT} + 352.695 \times RC_{BLI} + 63510 \quad (1)$$

$$APT = 1.4633 \times RC_{BAI} - 13.9032 \times RC_{NT} + 278.565 \times RC_{BLI} + 4372.6 \quad (2)$$

To determine the appropriateness of regression coefficients and goodness of regression model, various validation parameters are available [15–17].

Table 5 Regression parameters analysis

Performance parameter	Regression parameters					
	R-squared	Adjusted R-squared	F-statistic	Prob (F-statistic)	AIC	BIC
ASD	0.740	0.685	6.180	0.00169	890.4	897.2
APT	0.717	0.669	8.592	0.000196	703.8	710.5

Table 6 Interpretation and significance of various parameters involved in Regression analysis

Parameter	Interpretation
P-value	A predictor is significant if its P-value is less than 0.05
R-squared	Specifies closeness of fitted regression line with data. Ranges from 0 to 1. Higher value represents better fit
Adjusted R-squared	A bit lesser value than R-Squared and represents more accurate interpretation of goodness of regression model
F-Statistic and Prob (F-Statistic)	F-Test assesses overall significance of all coefficients collectively
AIC (Aiaike Information Criteria) and BIC (Bayesian Information criterion)	AIC assess multiple possible models with different set of variables. Lower AIC is considered better. BIC is similar to AIC but is more prominent in case of small sample size

Table 7 Optimized performance analysis

Batch interval 1000 ms			Batch interval 2000 ms		
Block Interval for stable performance	#Blocks = Batch interval/block interval	#Threads	Block Interval for stable performance	#Blocks = Batch interval/block interval	#Threads
350	2.85 ≈ 3	2	450	4.44 ≈ 5	2
400	**2.50 ≈ 3**	**4**	**500**	**4.00 ≈ 4**	**4**
400	2.50 ≈ 3	6	500	4.00 ≈ 4	6
400	2.50 ≈ 3	8	500	4.00 ≈ 4	8

Significance of various validation parameters of regression analysis is mentioned in Table 6.

The analysis shown in Table 3 depicts that the optimum performance of a streaming application stabilizes at a certain block interval, no matter if number of threads is increased further. To validate this behavior, we performed an analysis to find minimum a number of threads sufficient to saturate the block interval for stable performance of the streaming application in Spark. Table 7 lists calculations made for this analysis.

As we explored the relationship between the number of blocks emitted during a batch interval and number of threads used for processing, it is found that the optimum stable performance of a streaming application is stabilized as the number of threads becomes equal or greater than the number of emitted blocks during a batch interval. Minimum number of threads required to stabilize the streaming application's performance for a certain batch interval is highlighted in Table 7.

4 Related Work

The need for large-scale distributed data processing has given birth to the development of a wide variety of cluster computing systems [18–20]. A lot of applications need real-time processing [21, 22] and in-stream processing capabilities [23–25]. These fast growing in-motion has led to emergence of stream processing engines (SPEs) such as Yahoo's S4 [3], Twitter's Storm [4], Apache Spark [7, 8], Mill Wheel, [26], framework [27]. Among these, Apache Spark is being widely used over several hundreds of production deployments as it offers variety of workloads like batch jobs, data streams, graph analytics and SQL [9, 28–30], online aggregation [31], large-scale neuroscience [32], and genomic data processing [33]. Prasad [34] and Ditzler [35] discussed predictive modeling techniques for streaming data. Zliobaite et al. [36] worked on concept drift during online processing. Collaborative data stream mining is discussed by Gaber et al. [37]. Feng et al. [38] specified system for merging online transaction processing with stream processing for high transaction throughput. Jiang et al. [39] studied tuning of Mapreduce task under various levels of parallelism. Armbrust et al. [40] focused on memory and network layer management to enhance the performance of Spark. Cong et al. [24] mapped streaming applications onto Field Programming Gate Arrays to optimize the parallel computation. Davidson et al. [41] tried to resolve the inefficiency in shuffle phase of Spark for speed ups. Amos et al. [42] studied the impact of spark tuning options on Spark's performance to get an improvement of up to 5 times.

Several research works have been done in previous years regarding streaming applications and its performance enhancements in Spark. This work aims to explore research in different dimension by evaluating the impact of control parameters on performance of streaming applications and provide a prediction model for the scheduling delay and processing time to predict these parameters in advance before deploying application. Moreover, optimal performance analysis is done with respect to control parameters. In this way, this work makes significant contributions in research regarding performance prediction and optimization of streaming applications on Spark.

5 Conclusion and Future Work

Experimental analyses in this work explore streaming application performance with respect to various configurable streaming parameters in Spark. It also explains that how the tuning of these parameters affect the performance. A relationship is established among the performance and streaming parameters using a linear regression model, which enables the performance prediction of streaming application with respect to a specific combination of tuning parameters before actual deployment of the application. Finally, an analysis is done to determine the selection of minimum number of threads for parallel data processing which results in optimum performance while sustaining stability of application. In this work, we focused only on the data processing parallelism and batch and block interval parameters. In future, the behavior of the streaming application can explored in light of other parameters too such as memory parameter, data receiving parallelism, etc. Also, these analyses can be extended to analyze the performance with respect to various cluster size for computing.

References

1. Abadi, D. J., Carney, D., Çetintemel, U., Cherniack, M., Convey, C., Lee, S., Zdonik, S.: Aurora: a new model and architecture for data stream management. The VLDB Journal—The International Journal on Very Large Data Bases, 12, 2 (2003) 120–139.
2. Bifet, A., Holmes, G., Kirkby, R., Pfahringer, B.: MOA: DATA STREAM MINING - A Practical Approach. The University of Waikato, (2011).
3. Neumeyer, L., Robbins, B., Nair, A., Kesari, A.: S4: Distributed stream computing platform. In: IEEE International Conference on Data Mining Workshops (ICDMW' 10), pp. 170–177, IEEE Press, Washington DC, USA (2010).
4. Leibiusky, J., Eisbruch, G., Simonassi, D.: Getting Started with Storm-Continuous Streaming Computation with Twitter's Cluster Technology. O'Reilly, (2012).
5. Murdopo, A., Severien, A., Morales, G.D.F., and Bifet, A.: SAMOA: Developer's Guide. Yahoo Labs, (2013).
6. Prasad, B. R., Agarwal, S.: Handling Big Data Stream Analytics using SAMOA Framework-A Practical Experience. Int. J. Database Theory & Application, 7, 4 (2014).
7. Zaharia, M., Chowdhury, M., Franklin, M. J., Shenker, S., Stoica, I.: Spark: cluster computing with working sets. In: 2nd USENIX Conference on Hot Topics in Cloud Computing (HotCloud'10), pp. 10–10, Berkeley, USA: USENIX Association (2010).
8. Hamstra, M., Karau, H., Zaharia, M., Konwinski, A., Wendell, P.: Learning Spark: Lightning-Fast Big Data Analysis. O'Reilly Media, Inc., (2015).
9. Zaharia, M., Chowdhury, M., Das, T., Dave, A., Ma, J., McCauley, M., Franklin, M. J., Shenker, S., Stoica, I.: Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing. In: 9th USENIX Conference on Networked Systems Design and Implementation (NSDI' 12), pp. 2–2, USENIX Association (2012).
10. Prasad, B. R., Agarwal, S.: High speed streaming data analysis of web generated log streams. In: 10th IEEE International Conference on Industrial and Information Systems (ICIIS' 15), pp. 413–418, IEEE-Press, Peradeniya, Sri Lanka (2015).
11. Spark Streaming Programming Guide. https://spark.apache.org/docs/1.2.0/streaming-programming-guide.html#level-of-parallelism-in-data-receiving.

12. Spark Configuration. http://spark.apache.org/docs/latest/configuration.html#scheduling.
13. NASA Dataset source, http://www.ita.ee.lbl.gov/html/contrib/NASA-HTTP.html.
14. Chatterjee, S., Hadi, A. S.: Regression analysis by example. John Wiley & Sons, (2015).
15. Draper, N. R., Smith, H., Pownell, E. Applied regression analysis. John Wiley & Sons, New York (2014).
16. Shirley, M. W., Patel, N.: Estimating Beta: Interpreting Regression Statistics. Cost of Capital: Applications and Examples, (2014) 234–242.
17. Ashenfelter, O., Levine, P. B., Zimmerman, D. J.: Statistics and econometrics: methods and applications. John Wiley & Sons, New York (2003).
18. Dean, J., Ghemawat, S.: MapReduce: simplified data processing on large clusters. Communications of the ACM, 51, 1 (2008) 107–113.
19. Isard, M., et al.: Dryad: distributed data-parallel programs from sequential building blocks. ACM SIGOPS Operating Systems Review 41, 3 (2007) 59–72.
20. Malewicz, G., et al.: Pregel: a system for large-scale graph processing. In: ACM SIGMOD International Conference on Management of data, pp. 135–146, ACM, (2010).
21. M. Stonebraker, M., Çetintemel, U., Zdonik, S.: The 8 requirements of real-time stream processing. ACM SIGMOD Record, 34, 4 (2005) 42–47.
22. Barlow, M.: Real-time big data analytics: emerging architecture. O'Reilly Media, Inc., 2013.
23. Cugola, G., Margara, A.: Processing flows of information: From data stream to complex event processing. ACM Computing Surveys (CSUR), 44, 3 (2012) 15:1–62.
24. Cong, J., Huang, M., Zhang, P.: Combining computation and communication optimizations in system synthesis for streaming applications. In: ACM/SIGDA International Symposium on Field-Programmable Gate Array, pp. 213–222, ACM, (2014).
25. Kim, G. H., Trimi, S., Chung, J. H.: Big-data applications in the government sector. Communications of the ACM, 57, 3 (2014) 78–85.
26. Broekema, P. C., Boonstra, A. J., Cabezas, V. C., Engbersen, T., Holties, H., Jelitto, J., Ronald P. L. Offrein, B. J.: DOME: towards the ASTRON & IBM center for exascale technology. In: Workshop on High-Performance Computing for Astronomy Date, pp. 1–4, ACM, (2012).
27. Akidau, T., et al.: MillWheel: MillWheel: Fault-Tolerant Stream Processing at Internet Scale. In: VLDB Endowment, 6, 11, pp. 1033–1044, (2013).
28. Armbrust, M., et al.: Spark SQL: Relational data processing in Spark. In: ACM SIGMOD International Conference on Management of Data, pp. 1383–1394, ACM (2015).
29. Zaharia, M., Das, T., Li, H., Hunter, T., Shenker, S., Stoica, I.: Discretized streams: Fault-tolerant streaming computation at scale. In: 24th ACM Symposium on Operating Systems Principles, pp. 423–438, ACM, (2013).
30. Gonzalez, J. E., et al.: Graphx: Graph processing in a distributed dataflow framework. In: Proceedings of OSDI, pp. 599–613, (2014).
31. Zeng, K., Agarwal, S., Dave, A., Armbrust, M., Stoica, I.: G-OLA: Generalized On-Line Aggregation for Interactive Analysis on Big Data. In: ACM SIGMOD International Conference on Management of Data, pp. 913–918, ACM, (2015).
32. Freeman, J., Vladimirov, N., Kawashima, T., Mu, Y., Sofroniew, N. J., Bennett, D. V., Rosen, J., Yang, C. T., Looger, L. L., Ahrens, M. B.: Mapping brain activity at scale with cluster computing. Nature methods, 11, 9 (2014) 941–950.
33. Nothaft, F. A., et al.: A. Rethinking data-intensive science using scalable analytics systems. In: Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, pp. 631–646, ACM, (2015).
34. Prasad, B. R., Agarwal, S.: Critical parameter analysis of Vertical Hoeffding Tree for optimized performance using SAMOA. Int. J. Mach. Learn. and Cybernetics, (2016) 1–14.
35. Ditzler, G., Polikar, R.: Semi-supervised learning in nonstationary environments. In: IEEE International Joint Conference on Neural Networks, pp. 2741–2748, IEEE Press, (2011).
36. Zliobaite, I., et al.: Next challenges for adaptive learning systems. ACM SIGKDD Explorations Newsletter, 14, 1 (2012) 48–55.

37. Gaber, M. M., Gama, J., Krishnaswamy, S., Gomes, J. B., Stahl, F. Data stream mining in ubiquitous environments: state-of-the-art and current directions. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 4, 2 (2014) 116–138.
38. Feng, Y., Shen, X., Tian, J., Zhao, D., Wang, D., Zou, L.: S-store: An engine for large rdf graph integrating spatial information. In: 18th International Conference on Database Systems for Advanced Applications (DASFAA' 13), pp. 33, Springer, Wuhan, China (2013).
39. Jiang, D., Ooi, B. C., Shi, L., Wu, S.: The performance of mapreduce: An in-depth study. In: VLDB Endowment, 3, 1–2, pp. 472-483, VLDB, (2010).
40. Armbrust, M., Das, T., Davidson, A., Ghodsi, A., Or, A., Rosen, J., Stoica, I., Wendell, P., Xin, R., and Zaharia, M.: Scaling Spark in the Real World: Performance and Usability. In: VLDB Endowment, 8, 12 (2015).
41. Davidson, A., Or, A.: Optimizing Shuffle Performance in Spark. Technical Report, Berkeley, University of California, 2013.
42. Amos, B., Tompkins, D.: Performance study of Spindle, a web analytics query engine implemented in Spark. In: 6th IEEE International Conference on Cloud Computing Technology and Science (CloudCom' 14), pp. 505–510, IEEE, (2014).

A Browser-Based Distributed Framework for Content Sharing and Student Collaboration

Shikhar Vashishth, Yash Sinha and K. Haribabu

Abstract The utilization of the networks in education system has become increasingly widespread in recent years. WebRTC has been one of the hottest topics recently when it comes to Web technologies for distributed systems as it enables peer-to-peer (P2P) connectivity between machines with higher reliability and better scalability without the overhead of resource management. In this paper, we propose a browser based, asynchronous framework of a P2P network using distributed, lookup protocol (Chord), NodeJS and RTCDataChannel; which is scalable and lightweight. The design combines the advantages of P2P networks for better and sophisticated education delivery. The framework will facilitate students to share course content and discuss with fellow students without requiring any centralized infrastructure support.

Keywords Chord · Peer-to-peer · Nodejs · Webrtc · RTCDataChannel · Decentralized distributed systems

1 Introduction

Several technologies like computing, database systems, networking, and web technology play an important role in the field of education technology. In the field of networking, several advancements have led to boom in online education. Beginning with

The original version of this chapter was revised: Author name has been updated. The erratum to the chapter is available at 10.1007/978-981-10-3376-6_61

S. Vashishth (✉) · Y. Sinha · K. Haribabu
Department of Computer Science and Information Systems, Birla Institute of Technology
and Science, Pilani, India
e-mail: f2012436@pilani.bits-pilani.ac.in

Y. Sinha
e-mail: f2012365@pilani.bits-pilani.ac.in

K. Haribabu
e-mail: khari@pilani.bits-pilani.ac.in

P.K. Sa et al. (eds.), *Progress in Intelligent Computing Techniques: Theory, Practice, and Applications*, Advances in Intelligent Systems
and Computing 719, DOI 10.1007/978-981-10-3376-6_12

traditional centralized systems, the trend has moved on to decentralized distributed systems. The use of such distributed systems promises several advantages, such as higher reliability and better scalability. The peer-to-peer technology is superior to traditional client–server model because it does not require setting up a third-party server which makes it much more economical and practical in many situations. The peer-to-peer architecture offers the promise of harnessing the vast number of nodes connected to the network. Other significant features are redundant storage, permanence, efficient data location, anonymity, search, authentication, and hierarchical naming. Thus, this technology can be highly beneficial in supporting teachers and students to collaborate and share information together within a community.

The contribution of this paper is design and successful implementation of a browser based, asynchronous framework of a P2P network using distributed, lookup protocol (Chord [1]), NodeJS, and RTCDataChannel. We have chosen Chord as it provides efficient lookup in a dynamic peer-to-peer system with frequent node arrivals and departures. Additional features can be layered on the top of the framework based on the practical usage to gain robustness and scalability. We have also illustrated some of the cases wherein this framework can be beneficial.

Current frameworks are not very suitable for web applications because the web browsers are generally single threaded. We enlist the reasons in a subsequent section and advocate the need for a framework based on event-driven programming paradigm and asynchronous calls for a web browser. We also improve the fault tolerance by making dependence on the bootstrap server weaker, which is a point of network failure. Therefore, this framework helps P2P technology to enter in more areas.

We explain the implementation details of the framework (Fig. 1) in the rest of the paper. Section 2 compares this asynchronous implementation to related work. Section 3 presents the reasons for this framework. Section 4 of the paper describes the base Chord protocol, notations used, the bootstrap server, join operations of new peers, data structures used, and handling of RTCDataChannel connections between peers. Section 5 presents the implementation details of the framework. Section 6 highlights certain ways in which this work can be used to enhance the learning experience in a community. Finally, conclude in Sect. 7.

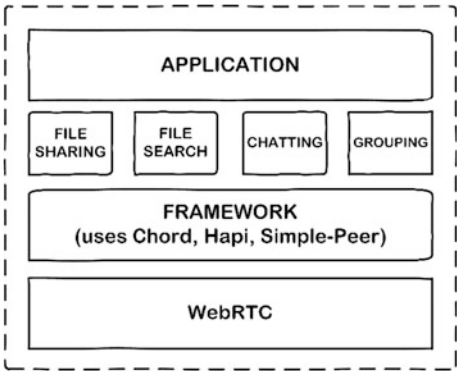


Fig. 1 Layers in the designed framework

2 Related Work

CHEWBACCA (CHord, Enhanced With Basic Algorithm Corrections, and Concurrent Activation) [2] is a P2P network framework in Java using sockets-based messaging. Although, this framework uses synchronous calls, it is not suitable for event-driven, non-blocking systems like web browsers. WebRTC-Chord is an asynchronous implementation of Chord protocol on Node Package Manager [3]. Although made for web browser, WebRTC-Chord is not as efficient because it does not keep bootstrapping server lightweight. The server stores all information about the peers and helps them to update their entries as the network changes dynamically over time. Thus, it is not as scalable as this framework. This framework reduces the involvement of bootstrap server only up to the initial phase of joining network for a peer. Establishing connection with other peers later does not involve bootstrap server at all. joonion-jchord [4] is simple implementation of Chord protocol written in Java is implemented for single virtual machine not for multiple systems.

3 Need for Framework

3.1 Web Browsers Incompatible with Synchronous Calls

Current frameworks of Chord are in C++ and Java and they use synchronous calls and the socket programming APIs. Processes, facilitated by the kernel, can wait for the synchronous calls to return, by getting suspended. Also threads can be used for foreground and background tasks. But synchronous calls in a web browser framework are discouraged for three reasons. First, browsers are built using event-driven programming paradigm, where asynchronous calls are suitable. Second, spinning or busy waiting locks the browser and other processes start crawling [5]. Thirdly, most browsers are single threaded and do not do anything on screen while Javascript code is running [6]. This hampers user experience badly. So, a need for an asynchronous framework was felt, designed and implemented.

3.2 Weakly Connected Bootstrap Server

There are frameworks, wherein the bootstrap server is actively involved to connect new peers and stabilize the network. The server stores all information about the peers and helps them to update their entries as the network changes dynamically over time [3]. Thus, there is a point of failure in the network, if the bootstrap server fails, new peers cannot join and the network cannot be stabilized. New RTCDataChannel connections are dependent on server to facilitate handshakes between the new peer and the network.

This framework, however, makes the dependence on the bootstrap server weaker by facilitating handshakes between new peer and the network via the boot peer and other peers in the network. The bootstrap server is only involved in assigning a new, unique id to a new peer and connecting it with a peer from the network chosen randomly. This also balances the load on peers to facilitate handshakes. But this requires more messages to be sent, forwarded, and accepted between peers in an asynchronous way. We have modified the function calls of the Chord protocol [1], to accommodate these changes.

4 Framework Design

4.1 Base Chord Protocol

Chord is protocol for peer-to-peer distributed hash table. It assigns each peer an m-bit identifier using base hash function such as SHA-1. Each peer maintains small amount of routing information that makes chord scalable by avoiding every peer to know about every other peer. In N-peer network each peer maintains information about only O(log N) other peers and a lookup requires O(log N) messages. Nodes and keys are arranged in an identifier circle that has at most 2^m peers, ranging from 0 to $2^m - 1$. The notations related are detailed in Table 1.

4.2 The Bootstrap Server

It allows new peer to join network by establishing its connection with one of the peer, which is already in network. For a new peer, peer becomes its boot peer whom it contacts for joining the network and establishing connection with other peers. Bootstrap server has no other role than to help peer in making connection with its boot peer, this helps to keep server light weight and makes the framework more scalable. It is implemented using hapi framework [7].

Table 1 Notations Used

Notation	Definition
Boot Peer	First peer (in network) with which a connection is established
Pred	Predecessor of peer in network
Succ	Successor of peer in network
Id	Unique Identifier of a peer
SuccPred	Predecessor of peer’s successor
CBF	Function to be called when remote call is complete
Signal	Signaling data of peer which is required to establish connection
Path	Stores the route which message took to reach destination

4.3 Peer Joins

Whenever a new peer joins a network, these invariants should be kept. It must have connection with at least one peer in network (boot peer). Each peer's successor points to its immediate successor correctly. Each key is stored in successor(k). Each peer's finger table should be correct. If finger table entries are not correct than query cost becomes O(n) instead of O(log(n)).

Following steps are involved for making a new peer join network. Initialize peer—assign it an identifier, initialize finger table, etc. Find the successor of new peer by querying bootstrap peer. Find the predecessor of new peer by asking peer's successor to send information about its predecessor. Notify other peers to update their successor, predecessors to maintain correctness of the framework. Calling fixFinger operation after regular interval to keep finger table entries up to date. Fix Finger operation involves calling findSucc operation for each finger table entry taking O(mlog(n)) time to update entire finger table.

4.4 Connection Objects: The Special Case

Chord protocol specification assumes that a peer can connect to a peer as long as it has its peer id. But in the real world, a connection needs to be established so that messages can be sent and received. We have used WebRTC's RTCDataChannel API to connect peers. The npm module Simple Peer [8] has been used to connect two machines. The initial handshake to connect to boot peer is facilitated by the bootstrap server, whereas subsequent handshakes to connect to other peers are handled by boot peer and other peers (as and when they get connected). The signals are sent from the source peer in the findSucc and findPred functions. The function calls hop over peers till a peer finally resolves the answer (some peer) of the given function call. This peer forwards the function call to the answer peer, which accepts the signal of the source peer. It then sends its own signal to the source peer, via the same path by which the function call arrived at it. Thus, the handshake is complete.

4.5 Implementation Details

Following data structures were used to store various information about the state of the network and connection objects. Response Table is an associative data structure maps message id to response received on making a remote call. Waiting Connection Table stores information about connection with which peer is currently in process of making connection. Connected Connection Table stores connection information about peer with which the peer already has established connection. Finger

Table contains m entries of peers with which peer has connection. It helps to avoid linear search. The ith entry of peer n will contain successor ((n + 2i − 1) mod 2 m). The first entry of finger table is the peer's immediate successor. Using finger table each query operation in network can be completed in O(log(n)).

4.6 Pseudo Code

joinNetwork. This is asynchronous implementation of joinNetwork operation which allows it to run in non-blocking, asynchronous frameworks such as NodeJS. This works same as synchronous version but is based on event-driven programming paradigm.

joinNetworkSynchronous():

```
n.pred = null; n.succ = b.findSucc(n.id);
n.succPred = n.succ.getPre(); n.stabilize();
if n.succPred == null: n.succ.stabilize();
else: n.succPred.stabilize();
```

joinNetworkAsynchronous(state, data):

```
case 0:  n.pred = null; msgId = respTable.new();
         n.initFindSucc(b.id, n.id, msgId, "next state")
case 1:  n.succ = respTable.get(msgId);
         msgId = respTable.new();
         n.initFindPred(n.succ, msgId, "next state")
case 2:  n.succPred = respTable.get(msgId);
         if n.succPred! = null AND n.succPred ∈ (n.Id,
         n.succ]
         n.succ = n.succPred;
         n.notifyPred(n.succ, n.id, msgId, "next state");
case 3:  if n.succPred == null:
         n.stabilize(n.succ, msgId, "next state");
else:    n.stabilize(n.succPred, msgId, "next state");
```

A peer executes **initFindSucc** and **initFindPred** to generate and attach signal before calling findSucc and findPred respectively. **findSucc** operation asks destPeer to tell successor of desired id. destPeer can contact other peers if it does not know the response to the query. When desired peer is found then signaling data of peer n is passed to that peer to allow it to form connection with it. The information about the order in which peers are contacted is stored in path variable of the call. It is used for returning final result and signaling data of successor of id to peer n. Once response reaches the peer, callBackFunc is invoked. **findPred** operation gives the immediate predecessor of given peer. When the predecessor of the desired peer is found, signaling data of peer is given to it which allows it to form connection with the calling peer. When peer calls **notifyPred** it asks destPeer to update the value of its predecessor to the given value of predecessor and CBF is called when this update request is complete. When a new peer joins the network then **stabilize** operation is called on given peer, its successor and predecessor to allow them to correct their predecessor and successor entries. This operation is essential for the framework to maintain its correctness. **fixFinger**. This operation is invoked by peers at regular intervals to update the entries of their finger tables which allows network to answer query in O(log(n)). If entries in finger table are not correct, query is resolved through successors which takes O(n) time to respond.

```
findSucc(destId, id, path, msgId, CBF):
if destId == n.id
if n.id == n.succ: exec(CBF);
elif id ∈ (n.id, n.succ]: n.succ.acceptSignal()
elif n.closePredFin(id) ==n.id:
n.findSucc(n.succ, id, path, msgId, CBF, signal);
else:n.findSucc(n.closePredFin(id),id,path,msgId,CBF);
else: path.append(n.id)
destId.findSucc(destId, id, path, msgId, CBF, signal);
```

```
findPred(destId, path, msgId, CBF, signal):
if destId == n.id: if n.id == n.succ: exec(CBF)
else: n.pred.acceptSignal()
else: destId.findPred(destId, path, msgId, CBF, signal)
```

```
notifyPred(destId, pred, path, msgId, CBF):
if destId == n.id: n.pred = pred;exec(CBF);
else: destId.notifyPred(destId, pred, path, msgId, CBF);
```

```
stabilize(destId, path, msgId, CBF):
if destId ==n.id:
n.succPred = n.initFindPred(n.succ, msgId, “next state”);
if n.succPred == null AND n.succPred ∈ (n.id, n.succ]:
n.succ = n.succPred
n.notifyPred(n.succ, n.id, msgId, next state”);
else: destId.stabilize(destId, path, msgId, CBF);
```

5 Networking Education

P2P technology is superior to traditional client–server model because it does not require setting up a third-party server which makes it much more economical and practical in many situations. Setting a third-party server involves expense on its security, maintenance, performance. It does not allow network to expand beyond a limit but such limitations are easily overcome by P2P technology because there is no central dependency on which whole network has to rely and security is also simplified because files' location are invisible to P2P peers and thus remains protected.

This framework does not require any kind of setup or installation on peer's system. Thus, web browsers can be part of the network, and very little expertise is required to create and join a network. We have illustrated some of the cases wherein this framework can be beneficial: Classroom Community, Off-campus Distance Education, Workflow management, Peer based Information Retrieval. Details have been skipped due to shortage of space.

6 Conclusion

Modern day browsers are single threaded applications, which cannot support synchronous calls to remote hosts. Busy waiting is not a feasible solution. We have explored the plausibility of a browser-based peer-to-peer network. This paper presents an asynchronous framework for P2P network built using distributed and lookup protocol called Chord, NodeJS and RTCDataChannel. Benefits of P2P networks include scalability, redundant storage, permanence, efficient data location, anonymity, search, authentication, and hierarchical naming. This allows for the framework to be easily used for promoting education in multiple scenarios. The framework enables institutions and students to share course content and discuss without overhead of resource management. The framework design includes scope for improvement. Security enhancements such as encryption can prevent a malicious peer to affect the network. The framework can be a part of browser integration such as plugin or extension.

References

1. Stoica, Ion, M. Robert, K. David, K. M. Frans and B. Hari, "Chord: A scalable peer-to-peer lookup service for internet applications," ACM SIGCOMM Computer Communication Review, vol. 31, no. 4, pp. 149–160, 2001.
2. Baker, Matthew, F. Russ, T. David and W. Adam, "Implementing a Distributed Peer to Peer File Sharing System using CHEWBACCA–CHord, Enhanced With Basic Algorithm Corrections and Concurrent Activation," 2003.
3. D. Dias, "webrtc-chord," [Online]. Available: http://www.npmjs.com/package/webrtc-chord.
4. "joonion-jchord," [Online]. Available: http://code.google.com/p/joonion-jchord/.
5. J. Wolter, "Javascript Madness: The Javascript Sleep Deficiency," [Online]. Available: http://unixpapa.com/js/sleep.html.
6. H.-G. Michna, "Sleep or wait in JavaScript| Windows Problem Solver," [Online]. Available: http://winhlp.com/node/633.
7. "hapi.js," [Online]. Available: http://hapijs.com/.
8. F. Aboukhadijeh, "simple-peer," [Online]. Available: http://www.npmjs.com/package/simple-peer.

Seed Point Selection Algorithm in Clustering of Image Data

Kuntal Chowdhury, Debasis Chaudhuri and Arup Kumar Pal

Abstract Massive amount of data are being collected in almost all sectors of life due to recent technological advancements. Various data mining tools including clustering is often applied on huge data sets in order to extract hidden and previously unknown information which can be helpful in future decision-making processes. Clustering is an unsupervised technique of data points which is separated into homogeneous groups. Seed point is an important feature of a clustering technique, which is called the core of the cluster and the performance of seed-based clustering technique depends on the choice of initial cluster center. The initial seed point selection is a challenging job due to formation of better cluster partition with rapidly convergence criteria. In the present research we have proposed the seed point selection algorithm applied on image data by taking the RGB features of color image as well as 2D data based on the maximization of Shannon's entropy with distance restriction criteria. Our seed point selection algorithm converges in a minimum number of steps for the formation of better clusters. We have applied our algorithm in different image data as well as discrete data and the results appear to be satisfactory. Also we have compared the result with other seed selection methods applied through K-Means algorithm for the comparative study of number of iterations and CPU time with the other clustering technique.

Keywords Data mining · Clustering · Seed point · Shannon's entropy · K-Means

K. Chowdhury (✉)
Department of Information Technology, DIT University, Dehradun 248001, India
e-mail: ikuntal09@gmail.com

D. Chaudhuri
DIC (DRDO), Panagarh Base, Muraripur, Bardhaman 713149, West Bengal, India
e-mail: deba_chaudhuri@yahoo.co.in

A.K. Pal
Department of Computer Science & Engineering, Indian Institute of Technology
(Indian School Of Mines), Dhanbad 826004, India
e-mail: arupkrpal@gmail.com

P.K. Sa et al. (eds.), *Progress in Intelligent Computing Techniques: Theory, Practice, and Applications*, Advances in Intelligent Systems and Computing 719, DOI 10.1007/978-981-10-3376-6_13

1 Introduction

Clustering is one of the noble and robust unsupervised technique upon which homogeneous data objects form a group to identify a particular class [1]. It is a process of partitioning a given data into homogeneous classes depending on certain characteristics. In the recent years, clustering analysis is considered to be the most useful technique for data mining applications. Clustering technique has widely been used in different fields of real life applications like, big data analytics [2], wireless sensor networks [3], intrusion detection [4], market segmentation [5], medicinal application [6], pattern recognition [7, 8], and genetics [9], etc. Partitional and Hierarchical clustering are the two ways of clustering algorithms [1, 10]. Hierarchical clustering continues through a series of partitions, and it is appropriate for information processing. Hierarchical clustering has also been applied in image segmentation by histogram thresholding [11]. On the other hand partitional clustering is to find the single partition rather than several as in hierarchical method. It is applicable on large datasets due to the formation of disjoint cluster. But its quality depends upon choice of number of output clusters and the choice of the initial seed points. It can be applied on whole dataset in order to find out global optimal or local optimal. *K*-means is very popular partitional clustering technique in terms of simplicity, robustness for the application in large real life data but with a disadvantage of prior knowledge of *K* (number of cluster) [12].

Researchers have attempted to know the optimal value of *K* through their own ideas in different literatures [13]. Global optimum results cannot be produced by *K*-Means algorithm due to its randomness in initial seed point selection phase. Seed-based algorithms are also appropriate for the cluster of spherical and ellipsoidal shape but for the cluster of arbitrary or elongated shape good results may not be obtained. To solve this cluster shape problem Chaudhuri et al. [14] have suggested the multi-seed concept where more than one seed may exist in a single cluster for the detection of shape and capture the appropriate cluster. Soft computing techniques have also been used for the detection of seed points in clustering by the use of Fuzzy logic [15].

Different researchers have proposed different hard decision algorithms regarding the initialization of *K*-means algorithm to achieve global optimum results in the different literatures mentioning the advantages and disadvantages represented in a comparative manner [7, 16–19]. Lu et al. [16] have suggested a weighted clustering technique to get the better seed point for the minimization of the number of iterations, computation time than the existing approaches. This weighted clustering problem is also capable of handling the influence of noise. Cao et al. [7] have also defined mathematical function based on rough set model for the detection of initial center point as well as border points. Bai et al. [18] have suggested density-based approach for cluster center initialization using the distance between the objects. This method is applicable to the categorical data having small number of clusters to produce good optimal solution. Reddy et al. [17] have also presented another approach for the selection of initial seed points using Vornoi diagram. Celebi et al.

[19] have suggested an efficient initialization approach in K-Means algorithm. Zahra et al. [20] have applied their proposed centroid selection in the design of recommender system. Density concept has also been used to find out the seed point as a highest density point. Astrahan [21] has taken the highest density point as an initial seed point and determined the other seed points depending on the distance criteria with decreasing density applied on speech data. Ball and Hall [22] suggested the mean value of the data set as the initial seed point with a distance criteria to obtain the other seed points. Chaudhuri et al. [23] have suggested the subset of seed points from a given superset using the cumulative density-based analysis. In their paper they have suggested the proper guidelines regarding the selection of the number of clusters and distance. Entropy minimization algorithm has been also used in clustering in different literatures.

In this paper, the proposed seed point detection algorithm for image data is on the basis of maximization of Shannon's entropy feature with distance restriction criteria. The proposed seed point selection algorithm converges in a minimum number of steps for the formation of better clusters. We have applied our algorithm in different image data as well as discrete data and the results appear to be satisfactory. Also we have compared the result with other seed selection algorithm for the comparative study of number of iterations in clustering technique. This paper is represented in the following manner. In Sect. 2 we have described our seed point selection algorithm. In Sect. 3 we have presented the comparative result of our algorithm with other clustering algorithms. In Sect. 4 we have concluded the future direction of our work that can be further made through the proposed seed selection approach.

2 Proposed Seed Point Selection Algorithm

The proposed seed point detection technique depends on the maximization of entropy-based objective function of multi-dimensional mutually independent variables with distance restriction criteria. Image processing operations on the bands of the multispectral or hyper-spectral images can be categorized into two types, either scalar image oriented or vector image oriented. In scalar image oriented operations on the bands of the images are processed in an independent manner. In vector image oriented the operations are mainly dependent on the vector nature of each pixel. Here we have taken the concept of scalar image oriented operation. The random intensity values of different bands in a particular pixel are independent due to the mutual independence on its wavelengths. Let R, G, and NIR are the three multispectral bands and the corresponding intensity values are represented as the triple (r, g, ir). Since the bands are mutually independent so the probability is defined as

$$P((R=r),(G=g),(NIR=ir)) \\ =P(R=r)P(G=g)P(NIR=ir) \tag{1}$$

Now $P(R=r)=\frac{n_r}{N}$, $P(G=g)=\frac{n_g}{N}$, and $P(NIR=ir)=\frac{n_{ir}}{N}$ where n_r, n_g, and n_{ir} are the number of points with gray value r, g, and ir, respectively. N is the total number of points in the data set.

Now the entropy of a multi-dimensional pixel in the data set is defined as

$$\begin{aligned} E= & -P\big((R=r_i),(G=g_i),(NIR=(ir)_i)\big)\log P\big((R=r_i),(G=g_i),(NIR=(ir)_i)\big) \\ = & -P(R=r_i)P(G=g_i)P(NIR=(ir)_i)[\log P(R=r_i)+\log P(G=g_i) \\ & +\log P(NIR=(ir)_i] \end{aligned} \tag{2}$$

Here we have demonstrated the proposed seed point selection algorithm.

Step 1: Find entropy (E) of each data using the above Eq. (2).

Step 2: Arrange the samples in decreasing order based on entropy values.

Step 3: Take the point whose entropy is maximum as the first seed point, say $(r_1, g_1, (ir)_1)$ and let S be the set of seed points.

Step 4: Take the point of second maximum entropy and calculate the distance between the considering point and the previous seed point. If the distance is less than T (threshold value) then do not consider the point in the seed point set S. Otherwise, the second maximum entropy point will be included as a second seed point in the seed point set S.

Step 5: Next consider the next highest entropy point and find all distances from the considering point and all other previous seed points from set S. Find the minimum distance, say $\min\_d$ and if $\min\_d < T$ then do not consider the point as another seed point. Otherwise considering point will be another seed point and update the set S accordingly.

Step 6: Repeat the Step 5 and stop if the number of residue points are very small.

Step 7: Stop.

3 Experimental Results

We have minutely analyzed multiple sets of real life image data for the determination of performance of the proposed seed point selection algorithm. Figure 1a shows 3-D multi-spectral training data set of different classes of total 30000 data. Total 3 seed points are detected using threshold $T = 170$. The corresponding clustered data by applying seed points selected by our proposed method through K-means clustering on data set is shown in Fig. 1b.

Similarly Fig. 2a shows 64516 3-D multi- spectral training data set of different classes. Total 4 seed points are detected using threshold $T = 110$. The corresponding

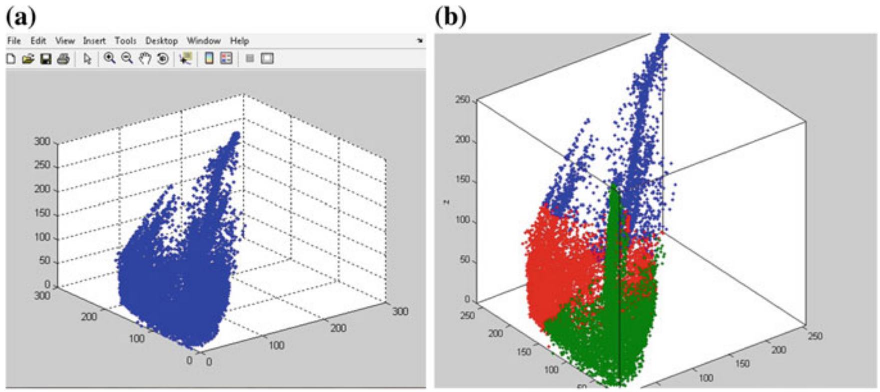


Fig. 1 **a** Original 3-D training data set and **b** clustered data set

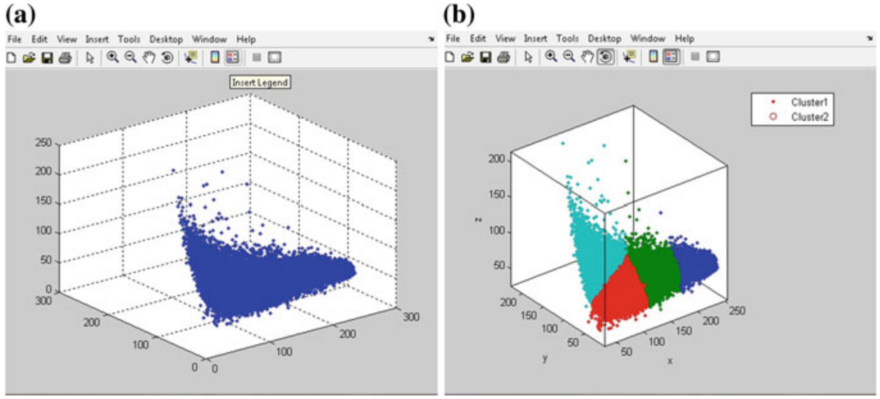


Fig. 2 **a** Original 3-D training data set and **b** clustered data set

clustered data by applying seed points selected by our method through *K*-means clustering on data set is shown in Fig. 2b.

In this paper, we have made the comparison of the performance of *K*-means clustering algorithm using the proposed seed point detection approach with the other classical clustering methodologies on different training data sets. Numerical experiments are conducted in MATLAB 7 in the Intel (R) Pentium 2.16 GHz hardware environment and Windows 7 with 4 GB RAM to show that our algorithm is better in terms of CPU time and number of iterations. We have also compared our proposed seed-based *K*-means clustering technique (PKM) results with other different clustering algorithms like Macqueen *K*-means (MKM), Fuzzy *C*-means (FCM), Expectation Maximization (EM) algorithm, Chaudhuri seed-based K-means clustering technique [23] (CKM) for different values of *K* and parameter *T*. Table 1 shows that both CPU time and number of iteration are much less by

Table 1 Comparison of the values of iteration and CPU time between the proposed and other techniques

Data	Size	Dimension	*K*	*T*	No. of iteration					CPU time (s)				
					MKM	FCM	EM	CKM	PKM	MKM	FCM	EM	CKM	PKM
Data 1	30000	3D	3	170	73	66	21	45	18	1.6	3.6	120.3	1.7	1.3
			4	150	67	100	25	38	22	0.8	5.6	234.1	0.6	0.4
			6	130	28	146	26	35	24	0.9	9.0	353.9	0.7	0.8
			9	110	59	167	37	49	31	1.3	7.1	501.9	1.9	0.9
Data 2	64516	3D	3	160	55	95	32	40	26	1.9	5.6	231.2	1.5	1.1
			4	110	62	76	45	48	29	2.1	5.4	425.5	1.85	1.4
Data 3	1378	2D	4	150	14	107	19	35	16	0.03	0.1	8.1	0.09	0.03
			7	150	33	107	16	32	13	0.1	0.3	11.0	0.14	0.06
			9	130	42	107	20	67	15	0.2	0.3	15.8	0.49	0.08
			11	110	37	43	18	16	14	0.2	0.6	22.76	0.09	0.07
Data 4	1019	2D	3	165	5	68	13	9	4	0.03	0.03	6.9	0.03	0.01
			6	110	23	123	21	18	15	0.06	0.31	3.24	0.09	0.05
			11	90	28	96	31	32	12	0.11	0.19	3.07	0.42	0.11

PKM than the other clustering technique, which means that the proposed seed point selection is more appropriate than the others.

4 Conclusion and Future Scope

We have presented an approach for detecting seed points applied on image data by taking RGB features of color image as well as 2D data based on the maximization of Shannon's entropy feature with distance restriction criteria. The proposed seed point selection algorithm converges in a minimum number of steps for the formation of better clusters. We have applied our algorithm in different image data and the results appear to be satisfactory. Also we have compared the result with other seed selection algorithm for the comparative study of number of iterations in clustering technique. This algorithm may be extended in the determination of the optimal number of clusters from unknown real data sets automatically with help of the entropy-based objective function. Also we may design better clustering technique using present seed point detection algorithm in future. This algorithm can also be applied on different class data for the development of efficient supervised classification scheme. We can also reduce the structural complexity of the image data for representation and also represent the whole image structure in compressed format using our seed point selection algorithm by the help of distance restriction criteria.

References

1. Jain, A. K., Dubes, R. C.: Algorithms for Clustering Data. Englewood Cliffs NJ: Prentice-Hall, (1988).
2. Singh, D., Reddy, C. K.: A survey on platforms for big data analytics. Journal of Big Data, Springer, 2(8), 1–20, doi:10.1186/s40537-014-0008-6, (2014).
3. Liu, Z., Zheng, Q., Xue, L., Guan, X.: A distributed energy efficient clustering algorithm with improved coverage in wireless sensor networks. Journal of Future Generation Computer System, 28(5), 780–790, (2012).
4. Wang, Q., Megalooikonomou, V.: A clustering Algorithm for intrusion detection. *In* Proc. of SPIE, 5812, 31–38, doi:10.1117/12.603567, (2005).
5. Kodabagi, M. M., Hanji, S. S., Hanji, S. V.: Application of enhanced clustering technique using similarity measure for market segmentation. CS&IT –CSCP-2014, 15–27, (2014).
6. Villmann, T., Albani, C.: Clustering of categoric data in medicine application of evolutionary algorithms. International Conference 7th Fuzzy Days on Computational Intelligence, Theory and Applications, 619–627, (2001).
7. Cao, F., Liang, J., Jiang, G.: An initialization for the K-Means algorithm using neighborhood model. Computers and Mathematics with Applications, **58**, 474–483, (2009).
8. Tou, J. T., Gonzales, R. C.: Pattern Recognition Principles. Addison-Wesley, (1974).
9. Bhattacharya, A., De, R. K.: Divisive correlation clustering algorithm (DCCA) for grouping of genes detecting varying patterns in expression profiles. Bioinformatics, **24**, 1359–1366, (2008).

10. Reddy, C. K., Vinazmuri, B.: A survey of partitional and hierarchical clustering algorithms. Data Clustering Algorithms and Applications, 87–110, (2013).
11. Arifin, A. Z., Asano, A.: Image segmentation by histogram thresholding using hierarchical cluster analysis. Pattern Recognition Letters, **27**(13), 1515–521, (2006).
12. Jain, A. K.: Data Clustering: 50 Years beyond K-Means. Pattern Recognition Letters, **31**(8), 651–666, (2010).
13. Chen, K., Li, L..: The best *K* for entropy based categorical data clustering. Proc. of International Conference on Scientific and Statistical Database Management (SSDBM), 253–262, (2005).
14. Chaudhuri, D., Chaudhuri, B. B.: A novel multi-seed nonhierarchical data clustering technique. IEEE Trans. on Systems, Man and Cybernetics – Part B: **27**(5), 871–877, (1997).
15. Pal, S.K., Paramanik, P. K.: Fuzzy measures in determining seed point in clustering. Pattern Recognition Letters, **4**, 159–164, (1986).
16. Lu, J. F., Tang, J. B., Tang, Z. M., Wang, J. Y.: Hierarchical initialization approach for K-Means clustering. Pattern Recognition Letters, **29**, 787–795, (2008).
17. Reddy, D., Jana, P. K.: Initialization for K-means clustering using Vornoi diagram. Procedia Technology 4, 395–400, (2012).
18. Bai, L., Liang, J., Dang, C., Cao, F.: A cluster centers initialization method for clustering categorical data. Expert Systems with Applications, **39**, 8022–8029, (2012).
19. Celebi, M. E., Hassan, A. K., Vela, P. A.: A comparative study of efficient initialization methods for the k-means clustering algorithm. Expert Systems with Applications, **40**, 200–210, (2013).
20. Zahra, S., Ghazanfar, M. A., Khalid, A., Naeem, U.: Novel centroid selection approaches for K-means-clustering based recommender systems. Information Sciences, **320**, 156–189, (2015).
21. Astrahan, M. M.: Speech analysis by clustering, or the hyperphoneme method. Stanford Artif. Intell. Proj. Memo. AIM-124, AD 09067, Stanford Univ., Stanford, CA, (1970).
22. Ball, G. H., Hall, D. J.: ISODATA: A novel method of data analysis and pattern classification. Tech. Rep. Stanford Res. Inst., Menlo Park, CA, (1965).
23. Chaudhuri, D., Murthy, C. A., Chaudhuri, B. B.: Finding a subset of representative points in a data set. IEEE Trans. on Systems, Man and Cybernetics, **24**(9), 1416–1424, (1994).

Comparative Analysis of AHP and Its Integrated Techniques Applied for Stock Index Ranking

H.S. Hota, Vineet Kumar Awasthi and Sanjay Kumar Singhai

Abstract Selection of stock index is a crucial task in financial decision-making process, especially when selection criterion is conflicting in nature. Multicriteria decision-making (MCDM) method like analytical hierarchy process (AHP) is one of the most widely used method, which may be utilized in the financial domain. This paper utilizes AHP and its integrated approaches using technique for order preference by similarity to ideal solution (TOPSIS) and simple additive weighting (SAW) for ranking of stock index. Three financial years data of six indices with six criteria are considered in the selection process. Experimental results reveals that S&P BSE SENSEX index is performing consistently well for all three financial years in case of all the techniques.

Keywords Multi Criteria Decision Making (MCDM) • Analytical Hierarchy Process (AHP) • Technique for Order Preference by Similarity to Ideal Solution (TOPSIS) • Simple Additive Weighting (SAW) • Stock index

1 Introduction

Ranking of stock index may be the requirement of financial experts, investors and other stakeholders associated in stock market for decision-making process. Selection of best index is a difficult process especially when their criterion is conflicting in nature. One of the important goals of financial investment is to find out the best

H.S. Hota (✉)
Bilaspur University, Bilaspur, CG, India
e-mail: profhota@gmail.com

V.K. Awasthi
Dr C V Raman University, Kota, CG, India
e-mail: vineet99kumar@gmail.com

S.K. Singhai
Government Engineering College, Raipur, CG, India
e-mail: ersanjaysinghai@gmail.com

P.K. Sa et al. (eds.), *Progress in Intelligent Computing Techniques: Theory, Practice, and Applications*, Advances in Intelligent Systems and Computing 719, DOI 10.1007/978-981-10-3376-6_14

possible solutions to maximize the returns and simultaneously to minimize the risk. Index selection problem with conflicting criteria can be understood as multicriteria decision making (MCDM). Analytical hierarchy process (AHP), technique for order preference by similarity to ideal solution (TOPSIS) method and simple additive weighting (SAW) are some very popular optimizing methods for ranking.

Author [1] has applied AHP technique to evaluate alternative fuel for the Greek road transport sector with seven alternatives in which cost and policy aspects are considered for a fuel selection. Nadali et al. [2] has used AHP and SAW methods for allocating the labels of credit customer after that data-mining algorithm has been applied. Many other authors have done lots of work in this area by applying various optimization techniques.

This paper utilizes AHP and its integrated approaches with SAW and TOPSIS for ranking of stock index. Three financial year data of 2011–2012, 2012–2013, and 2013–2014 of six indices with six criteria are considered in the selection process. Experimental results reveals that BSE 30 index is performing consistently well for all three financial years in case of all the techniques.

2 Financial Data and AHP with Its Integrated Approach

Financial data as index data and overall process of applying AHP and its integrated technique for ranking of stock index is explained in detail as follows:

2.1 Stock Index

Stock index data used for experimental purpose is downloaded from http://www.bseindia.com. Six popular indices as six alternatives of BSE (Bombay Stock Exchange) are considered, these are S&P BSE SENSEX (A1), S&P BSE MID CAP (A2), S&P BSE SMALL CAP (A3), S&P BSE 100 (A4), S&P BSE 200 (A5), and S&P BSE 500 (A6) with six criterion of each alternative namely High (Cr1), Low (Cr2), Close (Cr3), P/E ratio (Cr4), P/B ratio (Cr5), and Dividend (Cr6).

2.2 Ranking with AHP and Its Integrated Technique

One of the most popular MCDM methods is AHP method. This method is integrated with other MCDM methods like TOPSIS and SAW and applied by many researchers in various domains for the ranking of available alternatives. The methods are explained in detail as follows:

(i) **AHP**—AHP is proposed by Satty [3], in which multiple criteria are placed into a hierarchy along with goal at the top and alternatives at the bottom, method utilizes normalized object data which is prepared using some normalization techniques. A pair wise comparison matrix is then constructed with M attributes, that is, a square matrix $B_{M \times M}$ whose elements denote the comparative importance between attributes. In the pair wise comparison matrix geometric mean, relative normalized weight (W), consistency index (CI), and consistency ratio (CR) are calculated where CR should be less than 0.1. When the weights are consistent then composite performance score is calculated to find the rank of stock index.

(ii) **AHP-TOPSIS**—TOPSIS is another MCDM method proposed by Hwang et al. [4] which is also used as an alternative of AHP method. However, TOPSIS can be integrated with AHP where AHP calculates weights and finally weights are utilizes by TOPSIS to find out final weights of available alternatives. The important steps of TOPSIS [5] after obtaining weights through AHP are as follows:

Step 0: Input are the weights obtained through AHP.

Step 1: Obtain the decision matrix after using a numerical scale for intangibles.

Step 2: Calculate normalized decision matrix R, as bellow:

$$r_{ij} = x_{ij} / \left[\sum_{i=1}^{M} x_{ij}^2 \right]^{1/2} \tag{1}$$

Step 3: V as weighted decision matrix is calculated by multiplying each column of R by the corresponding weight.

Step 4: Obtain the positive ideal (A*) and the negative ideal (A^-) solutions from V as calculated in step 3.

Step 5: Separation measures S* and S^- are calculated for all the alternatives, for i = 1… m as follows:

$$S_i^* = \sqrt[2]{\sum_{j=1}^{n} (v_{ij} - v_j^*)^2} \tag{2}$$

$$S_i^- = \sqrt[2]{\sum_{j=1}^{n} (v_{ij} - v_j^-)^2} \tag{3}$$

Step 6: Relative closeness to the ideal solution for each alternative (C_i*, i = 1… m) will be

$$C_i^* = S_i / (S_i^* + S_i^-) \tag{4}$$

Step 7: As a last step rank is determine the by arranging the alternatives in the descending order of C_i*, i = 1… m.

(iii) **AHP-SAW**—AHP-SAW is another popular integrated method of MCDM [6] in which weight obtained through AHP is further utilized in SAW. In SAW, each attribute is given a weight where each alternative is assigned with regard to every attribute. Steps of SAW are as follows:

Step1: Obtain the decision matrix after using a numerical scale for intangibles.

Step2: Calculate normalized decision matrix, R (r_{ij}, i = 1… m: j = 1… n) using

$$r_{ij} = x_{ij} / x_j^*, \quad \text{if the jth criterion is a benefit criterion} \tag{5}$$

and

$$r_{ij} = x_j^- / x_{ij}, \quad \text{if the jth criterion is a cost criterion} \tag{6}$$

Step 3: Calculate weighted score for each alternative through multiplying each row of R by weight which may be obtained through AHP.

Step 4: Calculate the rank of each alternative based on the final score.

3 Stock Index Ranking Using AHP and Its Integrated Techniques

Earlier work of Hota et al. [7] is continued in this research work for the purpose of comparison with integrated approach of AHP, in which they have used AHP method and obtained weights of each criteria and rank of alternatives as shown in Tables 1 and 2, respectively, for the financial year data of 2013–2014. Rank of other two financial years is also obtained with similar way.

Further AHP-TOPSIS method is applied for the same indices using the normalized matrix as shown in Table 3, after normalization using Eq. 1 of the data of financial year 2013–2014, weighted decision matrix V is calculated using step 3 of AHP-TOPSIS method and presented in Table 4 with positive ideal solution (PIS) and negative ideal solution (NIS) as highlighted in bold and underlined letters, respectively. Separation measures of each alternative is obtained (Table 5) using Eqs. 2 and 3, respectively, for PIS and NIS. Finally the relative closeness value is calculated using Eq. 4 and rank of six stock indices are obtained and presented in Table 6.

Integrated AHP-SAW is also applied for the ranking of stock index, the normalized data for the six stock indices is calculated using Eqs. 5 and 6 and presented in Table 7. This method also use the weight of each criteria obtained through AHP as

Table 1 Weights of corresponding criteria calculated through AHP method [7]

Cr1	Cr2	Cr3	Cr4	Cr5	Cr6
0.421	0.083	0.226	0.130	0.083	0.057

Table 2 Obtained rank using AHP for the financial year 2013–14 [7]

Alternative	Weight value	Rank
A1	0.904	I
A2	0.344	V
A3	0.447	II
A4	0.380	IV
A5	0.247	VI
A6	0.418	III

Table 3 Normalized stock index data

Alternative	Criteria					
	High	Low	Close	P/E ratio	P/B ratio	Dividend
A1	0.833	0.844	0.832	0.278	0.548	0.389
A2	0.263	0.247	0.263	0.183	0.223	0.425
A3	0.262	0.246	0.263	0.829	0.208	0.456
A4	0.250	0.247	0.249	0.263	0.479	0.386
A5	0.099	0.098	0.099	0.261	0.465	0.391
A6	0.308	0.305	0.308	0.251	0.401	0.394

Table 4 Weighted decision matrix (V)

Alternative	Criteria					
	High	Low	Close	P/E ratio	P/B ratio	Dividend
A1	**0.351**	**0.07**	**0.188**	0.036	**0.045**	0.022
A2	0.111	0.021	0.059	0.023	0.018	0.024
A3	0.111	0.020	0.059	**0.108**	0.017	**0.026**
A4	0.105	0.021	0.056	0.034	0.039	0.022
A5	0.042	0.008	0.022	0.034	0.038	0.022
A6	0.129	0.025	0.069	0.032	0.033	0.022

Table 5 Separation measurers

	A1	A2	A3	A4	A5	A6
Si*	0.072	0.332	0.278	0.293	0.364	0.266
Si-	0.386	0.079	0.115	0.077	0.024	0.103

Table 6 Obtained rank using AHP-TOPSIS for the financial year 2013–14

Alternative	CI*(Relative closeness)	Rank
A1	0.842	I
A2	0.192	V
A3	0.292	II
A4	0.207	IV
A5	0.060	VI
A6	0.278	III

Table 7 Normalized stock index data applied with SAW

Alternative	Criteria					
	High	Low	Close	P/E ratio	P/B ratio	Dividend
A1	1	1	1	0.335	1	0.852
A2	0.315	0.293	0.316	0.221	0.406	0.932
A3	0.315	0.291	0.315	1	0.381	1
A4	0.299	0.293	0.299	0.317	0.874	0.847
A5	0.119	0.117	0.119	0.314	0.848	0.858
A6	0.370	0.117	0.371	0.302	0.730	0.863

Table 8 Obtained rank using AHP-SAW for the financial year 2013–14

Alternative	Weighted score	Rank
A1	0.905	I
A2	0.344	V
A3	0.447	II
A4	0.380	IV
A5	0.247	VI
A6	0.398	III

Table 9 Comparative rank of various techniques for the financial year 2013–14

Alternative	AHP	AHP-TOPSIS	AHP-SAW
A1	I	I	I
A2	V	V	V
A3	II	II	II
A4	IV	IV	IV
A5	VI	VI	VI
A6	III	III	III

mentioned in Table 1. Using step 2 of AHP-SAW, the weighted score is calculated which identifies the rank of each stock index. Table 8 presents the weighted score and rank of stock index using AHP-SAW method for the financial year 2013–14.

The comparative rank for financial year 2013–14 for all the techniques is depicted in Table 9. This table clearly shows that all the techniques are producing same rank for all the indices, hence there is no inconsistency among the techniques used for stock index ranking.

4 Comparative Analysis of Rank of Three Financial Year Stock Index

In order to find out the strength of a particular index, financial data of three financial years considered in this study are utilized and applied with three AHP related techniques as explained above and ranks are calculated as mentioned in Sects. 2

Table 10 Year wise and technique wise rank comparison of stock indices

Alternative	Financial year 2013–14			Financial year 2012–13			Financial year 2011–12		
	AHP	AHP-TOPSIS	AHP-SAW	AHP	AHP-TOPSIS	AHP-SAW	AHP	AHP-TOPSIS	AHP-SAW
A1	I	I	I	I	I	I	I	I	I
A2	V	V	V	IV	IV	IV	IV	III	IV
A3	II	II	II	III	III	III	III	II	II
A4	IV	IV	IV	V	V	V	V	IV	V
A5	VI	VI	VI	VI	VI	VI	VI	VI	VI
A6	III	III	III	II	II	II	II	V	II

and 3 and shown in Table 10. This table clearly reflects the strength of S&P BSE SENSEX with its consistent performance as I rank for all the financial years as well in case of all the techniques, while S&P BSE SMALL CAP or S&P BSE 500 may be the second preferable index for the investor's point of view as these indices are ranked either II or III using 2 out of three techniques, on the other hand S&P BSE MID CAP and S&P BSE 100 are obtaining IV or V rank; however, S&P BSE 200 index is consistently performing worst with lowest rank (VI rank).

5 Conclusion

Ranking of stock index is necessary for the investment of asset in the stock market and also for creating a best portfolio but ranking of the stock index is not possible when their criterion are conflicting in nature. MCDM-based techniques may overcome this problem since these techniques find out rank of available alternatives through some mathematical process. This paper explores three MCDM techniques namely AHP, AHP-TOPSIS, and AHP-SAW for stock index ranking of six indices and found consistent results in terms of rank with I rank as S&P BSE SENSEX.

References

1. Tsita, K. G., Pilavachi, P. A.: Evaluation of alternative fuels for the Greek road transport sector using the analytic hierarchy process. Energy Policy, 48, (2012) 677–686.
2. Nasoli, A. Pourdarab, S. And Nosrotabadi, H.E.: Class Labelling of Bank Credits Customers Using AHP and SAW for Credit Scoring with Data Mining Algorithm. International Journal of Computer Theory and Engineering, 4 (2012).
3. Saaty, T. L.: Decision Making with Dependence and Feedback: Analytic Network Process, RWS Publications. Pittsburgh (2001).
4. Hwang, C. L., and Yoon, K. S.: Multiple attribute decision making: Methods and Applications, Berlin. Springer-Verlag, (1981).
5. R. V. Rao: Decision Making in the Manufacturing Environment, Springer, (2010).
6. Fishburn, P. C.: Additive Utilities with Incomplete Product Set: Application to Priorities and Assignment. Operation Research Society of America, Baltimore (1967).
7. Hota, H.S., Sharma, D.K. and Awasthi, V.K.: Stack Portfolio Using MCDM Technique, Conference Proceeding of International conference on Advances in Management and Decision Science, (2015) pp 1–6.
8. Web sources, http://www.bseindia.com (last accessed on October 2015).

A Proposed What-Why-How (WWH) Learning Model for Students and Strengthening Learning Skills Through Computational Thinking

Rakesh Mohanty and Sudhansu Bala Das

Abstract In-spite of facilities and ambiances available in an organization, novice students often face challenges to develop a right attitude and framework for productive learning. Through a case study in our university, we explore some difficulties and challenges faced by the freshers to strengthen and enhance their learning skills. We focus on generating questions from trivial to nontrivial level in a systematic way to explore the learning patterns. We propose a learning model, which we popularly call as What-Why- How (WWH) model, for providing a framework to strengthen learning skills. Computational thinking will be a fundamental skill, which can be used by everyone in future to strengthen and enhance learning. It is a thought process that involves formulating problems so that solutions can be represented as computational steps and algorithms. In our work, we integrate the computational thinking approach in our proposed WWH model of learning and develop a novel framework to resolve some of the challenges associated with learning skills of freshers in educational institutions.

Keywords Computing education · Learning · Computational thinking

1 Introduction

Learning is an indispensable skill that can be strengthened by intellectual activities and mental training. Intellectual activities involve creativity, exploration, innovation, formulating questions, answering questions, problem solving, and critical thinking. Mental training involves developing curiosity, interest, patience, perseverance, practice, competitive spirit, self-motivation, determination, and self-confidence.

R. Mohanty (✉) · S. Bala Das
Veer Surendra Sai University of Technology, Burla, Sambalpur, Odisha, India
e-mail: rakesh.iitmphd@gmail.com
URL: http://www.vssut.ac.in/

S. Bala Das
e-mail: baladas.sudhansu@gmail.com

P.K. Sa et al. (eds.), *Progress in Intelligent Computing Techniques: Theory, Practice, and Applications*, Advances in Intelligent Systems
and Computing 719, DOI 10.1007/978-981-10-3376-6_15

Strengthening and enhancing of learning skills is an important practice for growth and progress in all spheres of life. In this paper, we propose a new learning model for novice first year students in educational institutions to strengthen their learning skills.

1.1 Question–Answer Based Learning

Learning has been well defined from different perspectives by various researchers in the literature. Learning is a process of acquiring new knowledge, skills, and values through activities, events, study, and experience. Learning process can be initiated by generating few basic questions. By answering these questions, we enhance our learning skill. Question–Answering (Q-A) has been an active learning method used to acquire knowledge and strengthen learning skills. Questions raised must be well defined and answers generated must be correct. What, Why, and How are the three basic questions raised for initiating the learning process in our proposed What-Why-How (WWH) learning model. The What questions deal with information, terminologies, definitions, concepts, action, and meaning. The Why questions represent logic, reasoning, structure, analysis, theory, and description. The How questions address construction, steps, methods, processes, models, algorithms, and skills. If we generate questions from trivial to nontrivial level, the learning can be systematic and progressive. Progressive learning creates a significant impact on learner's mind. According to our knowledge, finding a standard model of raising questions in a systematic and logical way for strengthening learning skills has not been addressed in the literature.

1.2 Our Approach for Question–Answer-Based Learning

We address the above challenging issue by classifying the questions raised into various levels as mentioned below. Questions raised can be divided into four classes such as level 0 (L_0), level 1 (L_1), level 2 (L_2) and level 3 (L_3) based on their complexity and difficulty in understanding. Level number indicates the difficulty and intellectual level of the questions raised. Here 0 represents basic informative questions and 1 represents simple and trivial questions. Moderate and semi-trivial questions are represented by 2 and 3 represents the difficult and non trivial questions. Here, L_0 questions address the basic aspects about the topic we learn. Similarly, L_1 questions are reasoning based, L_2 questions are process based, and L_3 questions are concept based. The higher the level number, the more complex the questions raised and more difficulty in the learning process. The four primitive questions which initiate the learning process are—(a) What to learn? (b) Why to learn? (c) How to learn? and (d) How to learn effectively? Here we can consider (a) as L_0, (b) as L_1, (c) as L_2 and (d) as L_3 questions.

2 Case Study

We conduct a case study by considering a sample of 45 students who are in their first year graduate and post graduate course in our university to explore the challenges faced by them. Our case study is classified into two types such as *activity based and object based* case study.

In activity-based case study, we consider various aspects of learning activity to raise What-Why-How questions through students. In object-based case study, we consider the term computer as an object to raise What-Why-How questions through students. Our method involves asking three questions from trivial to non trivial level with the two word phrases—What+learning, Why+learning, and How+learning in the activity based case study. Similarly, in the object based case study, we ask students to generate three questions with the two word phrases—What+Computer, Why+Computer, How+Computer. Based on the questions generated by a sample of 45 students, we classify the questions into three classes such as Level 1 (L_1), Level 2 (L_2), and Level 3 (L_3). Here we present some of the challenges faced by the students for strengthening their learning skills. We present and summarize the questions raised by the students in a systematic way using our WWH model as shown in Tables 1 and 2. L_1 questions are based on generic view about the term. L_2 questions are more specific details, strategies and concepts. L_3 questions are analytical. Through an interactive question–answer session with the above-mentioned sample of 45 students, we explore the following challenges faced by the students for strengthening their learning skills. The major learning challenges faced by the students as observed in our case study are lack of mental attributes for learning such as attitude, interest, purpose, objective, curiosity, passion, and perseverance for learning. Difficulty in understanding, memorizing, and visualizing during learning, irregular

Table 1 Activity-based case study

Activity-learning (What, Why, How Questions with levels)			
	L_1	L_2	L_3
What	What is learning?	What is collaborative learning?	What is effective learning?
	What is the purpose of learning?	What are the different learning skills?	What is the most important learning skill?
Why	Why students need to learn?	Why choice of learning model is important?	Why students should follow a learning model?
	Why learning skills should be strengthened?	Why attitude plays a major role in learning?	Why students should be motivated for learning?
How	How students learn?	How students can learn better?	How students learning can be effective?
	How to measure learning skills?	How to enhance learning skills?	How to design a learning model?

Table 2 Object-based case study

Object-computer (What, Why, How Questions with levels)			
	L_1	L_2	L_3
What	What is a computer?	What are the different components of a computer?	What is the architecture of a computer?
	What are the features of a computer?	What features can added in a computer?	What features can be removed from the computer?
Why	Why computer was invented?	Why computer gained its importance?	Why learning computer is important?
	Why computer can be used to solve problems?	Why computer solves problems in less time?	Why students are addicted to a computer?
How	How computer works?	How computer work faster?	How computer solves problems?
	How viruses affect computer?	How computer viruses are created?	How computer viruses can be removed?

and unstructured learning patterns are the other important challenges faced by students. To address the above-mentioned challenges we propose a learning model to strengthen learning skills.

3 Our Proposed What-Why-How Learning Model

Our proposed model has four components to strengthen the learning skills such as learning environment, learning attributes, levels of learning, and learning approaches as shown in Fig. 1. Learning environment can be intrinsic (internal psychological aspects) or extrinsic (external environments). The learning attributes can be either activity based or object based. The learning levels can be L_0, L_1, L_2, L_3. The learning approaches are the sequence of steps followed during learning. To strengthen the learning skills we need to ask What, Why, and How questions in all the components of our learning model.

4 Computational Thinking—What, Why, How

Papert [1] first used the term computational thinking (CT). But it was first brought to forefront of research Community by Jeannette Wing [2]. She described CT as the combined strengths of human and computer brains to solve problems and accomplish tasks.

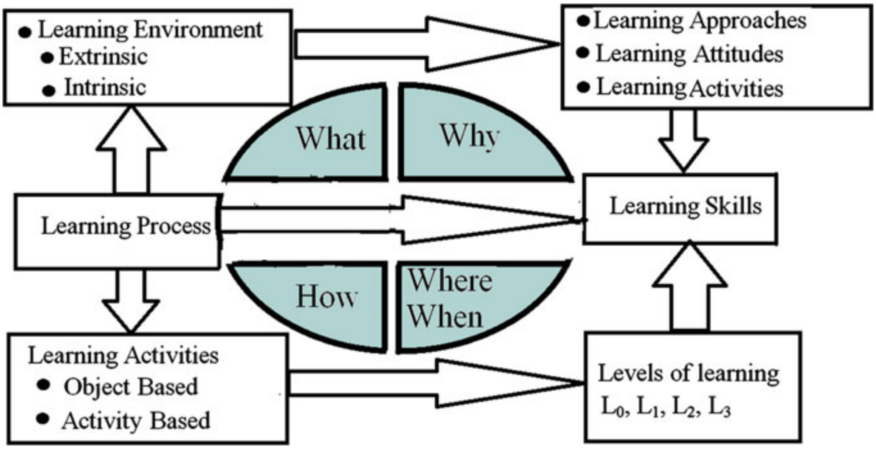


Fig. 1 Our proposed WWH learning model

Computational Thinking-What

CT is a problem-solving approach using different kinds of thinking such as analytical thinking, mathematical thinking, engineering thinking, scientific thinking, and critical thinking [2–4]. CT [5] helps us in organizing and analyzing data in logical manner, implementing and testing solutions and applying the solutions to other problems.

Computational Thinking-Why

We all use computers for different purposes such as sending email, browsing, playing game, or chatting. But, CT is a way to change our thinking process as it focuses on generating new ideas, describing hypotheses and theories. Without it, we will use incomplete and insufficient models, which can only develop faulty judgments about educational strategies. It makes use of well-known problem solving approaches such as trial and error finding, iteration, and guessing [5].

Computational Thinking-How

CT is based on five stages [6] such as decomposition, pattern recognition, pattern generalization, algorithm design, and evaluation. Decomposition is a way of breaking a problem into different subparts. Pattern recognition help us to observe different similarities, differences, and properties in data. Pattern generalization is an ability to remove unnecessary details, keeping only relevant data. Algorithm design help us to follow step-by-step process to solve different problem. Evaluation is about making judgments.

5 Computational Thinking-Based Learning

We can strengthen our learning skills by integrating CT in our WWH learning model by asking What-Why-How questions using various levels such as L_0, L_1, L_2, L_3 as shown in Fig. 2. During learning any concept, at first we need to identify the information needed to learn about that concept. Then, we break that concept into smaller sub problems and then we finally identify the specific information needed for learning that concept. After this we have to think about whether we have knowledge about concept like this before, and if so, how the new concept is different. We also look for patterns in this information which is helpful for us to consider how the information is structured, whether we have seen information organized like this before, and if so, how the new information is different. It is required to think about how the sub concepts are organized. Now, we have to use abstraction where we take only relevant information which is required for that concept. In algorithm design, we have to think about the steps from the initial information to the problem being solved and finally we have to see whether we have learn that concept or not.
Our CT-based learning model creates a motivation and drive among the students to foster the activities such as—active involvement, social participation, meaningful activities, relating new information to prior knowledge, being strategic, engaging in self-regulation and being reflective, restructuring prior knowledge, aiming toward understanding rather than memorization, taking time to practice, and Creating motivated learners.

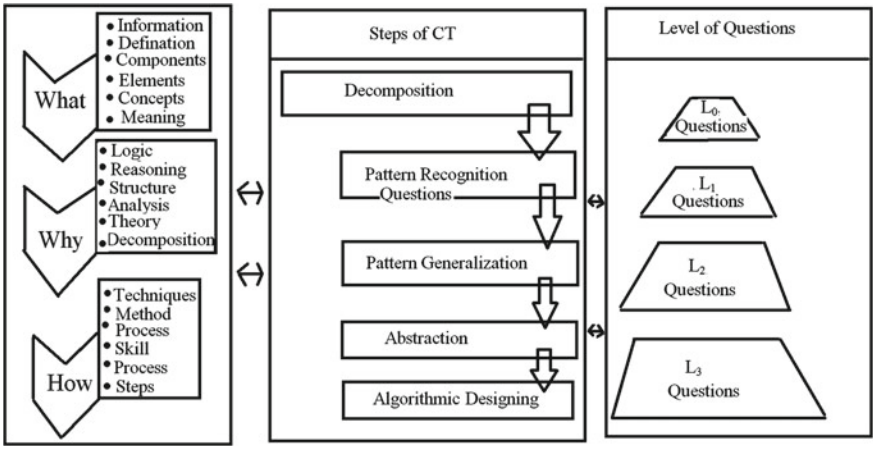


Fig. 2 Computational thinking-based learning model

6 Conclusion

Through a case study in our university, we explore some difficulties and challenges faced by the freshers to strengthen and enhance their learning skills. We focus on generating questions from trivial to non trivial level in a systematic way to strengthen the learning patterns of freshers. We proposed a What-Why-How (WWH) learning model to strengthen their learning skills. Then we integrated the computational thinking approach in our proposed What-Why-How (WWH) model to develop a novel framework to resolve some of the challenges associated with learning skills of freshers in educational institutions.

References

1. Seymour Papert, An exploration in the space of mathematics educations, *International Journal of Computers for Mathematical Learning*, 1(1), 95–123, 1996.
2. Jeannette M Wing, Computational Thinking, *Communications of the ACM*, 49(3): 33–35, 2006.
3. Jeannette M Wing, Computational Thinking and Thinking about Computing, *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 366(1881), 3717–3725, 2008.
4. Aman Yadav, Chris Mayfield, Ninger Zhou, Susanne Hambrusch and John T Korb, Computational Thinking in elementary and secondary teacher education, *ACM Transactions on Computing Education (TOCE)*, 14(1):5, 2014.
5. Valerie Barr and Chris Stephenson, Bringing Computational Thinking to K-12: what is Involved and what is the role of the computer science education community? *ACM Inroads*, 2(1): 48–54, 2011.
6. Elaine Kao, Exploring Computational Thinking at Google, *CSTA* Voice, 7(2), p 6, 2011.

Band Power Tuning of Primary Motor Cortex EEG for Continuous Bimanual Movements

Manikumar Tellamekala and Shaik Mohammad Rafi

Abstract Comprehension of natural intelligent systems' capability to exhibit bimanual coordination facilitates the process of building systems with better coordination dynamics. Dark side of the bimanual coordination, 'Bimanual interference', is that execution of continuous bimanual movements is heavily constrained by spatiotemporal coupling. Ability of callosotomy patients to draw circle and square patterns with two different hands simultaneously with perfect uncoupling after surgical removal of Carpus callosum, makes the point clear that Carpus callosum plays key role in bimanual interference. This paper introduces a new viewpoint of this phenomenon. While the right-handed subjects were drawing asymmetric clockwise and anticlockwise circles and symmetric circle-square patterns, neural activity of Primary motor cortex, which controls movement execution, is recorded. Here major emphasis is placed on how different frequency band powers of EEG signals from primary motor cortex are altered with respect to different continuous bimanual movements. Results in this study demonstrate the essence of understanding feedback loop connections between Corpus callosum and primary motor cortex.

Keywords Bimanual interference · Corpus callosum · Corticospinal pathways · EEG frequency band powers

1 Introduction

Coordination among multiple robotic arms in order to successfully accomplish a given task is put under limelight in recent days and is a prominent research trend. This process involves great deal of laborious and complicated design structures. For

M. Tellamekala · S.M. Rafi (✉)
Department of Electronics and Communication Engineering,
Rajiv Gandhi University of Knowledge Technologies, R. K. Valley, India
e-mail: rafi@rguktrkv.ac.in

M. Tellamekala
e-mail: manikumarrkv@gmail.com

P.K. Sa et al. (eds.), *Progress in Intelligent Computing Techniques: Theory, Practice, and Applications*, Advances in Intelligent Systems
and Computing 719, DOI 10.1007/978-981-10-3376-6_16

instance, realising human-like walking robots is not an easy job no matter how accurate mathematical modelling of respective mechanical systems is performed. This is solely because of the lack of knowledge regarding the coordination controller of these mechanical components. By closely observing, one admits the fact that natural intelligent systems perform the same tasks with utmost ease and comfort, popularly known as 'Bimanual coordination', [1] communication between the limbs to coordinate with each other to make a movement.

Understanding bimanual coordination of human hands [2] provides an opportunity to understand the precise dynamics of the coordination controller. At times, crosstalk between two hands may result in the behaviour, which oppositely oriented towards the prescribed goal. This effect is called as, 'Bimanual Interference' [3], dark side of the bimanual coordination. But in terms of the functionality, underlying controller properties and behaviour are same in both bimanual interference and coordination. So being knowledgeable about bimanual interference helps to understand bimanual coordination, which in turn leads to the successful emulation of the coordination controller architecture in artificial systems.

To put bimanual interference simply, this is the effect that prevents one to draw a square with one hand and a circle with another hand simultaneously [4]. Bimanual interference imposes strong constraints on the ability to make movements with limbs in both spatial and temporal domains [5]. Sub cortical neural structures are responsible for temporal coupling whereas Carpus callosum, i.e. communication channel between left and right cerebral hemispheres cause to spatial coupling [6]. Consider a task of drawing circles with left and right index fingers simultaneously, an extensively employed task to gain insights into the dynamics of spatiotemporal constraints on continuous bimanual movements. One can perform this task in two modes, symmetric and asymmetric. While drawing a circle moving one index finger in clockwise direction and another index finger in anticlockwise direction, creates a symmetric movement with body-midline reference. In asymmetric mode condition, two index fingers move in the same direction either clockwise or anticlockwise.

The stability of asymmetric mode circle drawing movements get deteriorated as the circle drawing rate goes high and gradually a transition from asymmetric mode to symmetric mode takes place. There exist two popular hypotheses to explain this deviation from asymmetric mode to symmetric mode while performing bimanual movements. The first hypothesis states that in symmetric mode, homologous muscles get activated by Ipsilateral descending corticospinal pathways and thus create signals in congruent fashion [7]. In case of asymmetric movements, non-homologous muscles get activated where in conflicts arise between crossed and uncrossed corticospinal pathways. This conflict leads to the deviation of the less dominant hand towards the symmetric motion.

The second hypothesis argues that deviation takes place due to the reinforcement of symmetric-drawing movements by corpus callosum, parietal, premotor and primary motor cortical regions [8]. Surgical removal of the carpus callosum in callosotomy patients with severe epilepsy, engendered to an important finding. This removal allowed patients to perform continuous bimanual moments with absolute spatial and temporal uncoupling [9]. Hence, the role of the carpus callosum in producing bimanual interference during dual motor tasks is prominent.

Apart from the above-mentioned hypothetical models of bimanual movements, recent studies [10] that investigated neural correlates of bimanual reaching movements put forth a completely new viewpoint of this discussion. In these studies, functional magnetic resonance imaging (fMRI) of the subjects while they were performing symmetric and asymmetric movements cued either directly or indirectly displayed increased activity in cingulate gyrus and pre-supplementary motor area. If the same subjects perform Stroop-test, a well-known cognitive interference task, the same regions get activated. These results suggest that bimanual interferences are not just leaks of electrical patterns at low-level communication channels such as corticospinal pathways and corpus callosal connections but are closely linked to cortical level controlling blocks. Essence of this inference is that bimanual interference is associated with not only the way a hand is to be moved to draw a geometrical pattern but also which hand has to be moved to generate the same pattern.

With the aim to observe the changes in the behaviour of one of the cortical level key control blocks of bimanual movements, primary motor cortex, the work that is discussed in the following sections is on how delta, theta, alpha, and beta band powers of primary motor cortex are getting tuned with the circle drawing movements.

2 Materials and Methods

Posterior parietal cortex, premotor cortex, and primary motor cortex are the prominent controlling blocks in producing continuous bimanual movements [11]. Information flow among these blocks is displayed in Fig. 1. Planning of movements is organized by the posterior parietal region and execution of planned movements is performed by the primary motor cortex. Free flow of movement execution is highly constrained in bimanual movements with spatiotemporal coupling. So, the following experimental set up is employed to gain insights into the impact of bimanual interference tasks on the motor cortex band specific electrical activity. Noninvasive EEG is selected for imaging of primary motor cortex in this work. Olimex open-EEG setup with amplifier and active EEG electrodes with good signal to noise ration are chosen for this purpose. As per 10–20 EEG electrode placement system, active electrodes are placed on C3 and C4 positions.

Asymmetric circle drawing in both clockwise and anticlockwise directions and symmetric mode circle-square drawing are the bimanual movements, performed by the right-handed subjects. Intended geometrical patterns and produced geometrical patterns by the right-handed subjects, are as shown in Fig. 2. During the execution of these three-task EEG data is recorded from the subjects' primary motor cortex. Acquired EEG signals could be the linear combination of motor cortex scalp potential differences and noisy scalp currents due to volume conduction phenomena, from other active cortical regions at that point of time. The weight vector of

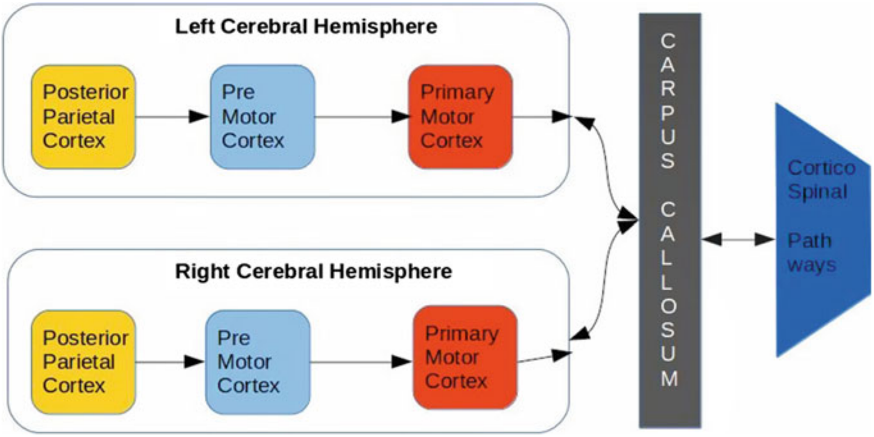


Fig. 1 Bimanual movement information flow through Posterior parietal cortex (planning of movements), Pre and primary motor cortices (movement execution), Carpus callosum (*left-right* hemispheres communication lines) to Ipsilateral descending pathways

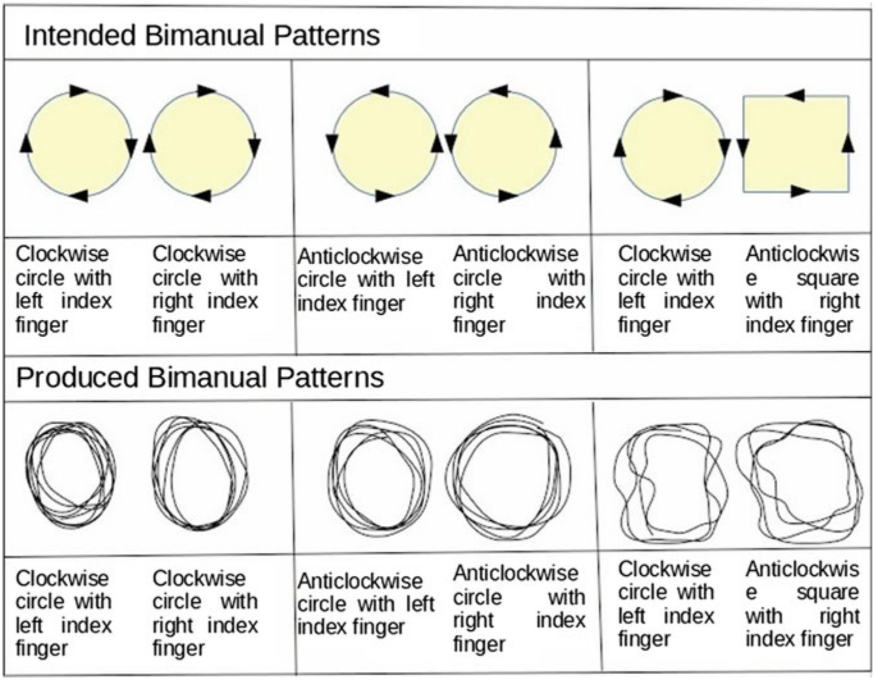

Intended Bimanual Patterns
Clockwise circle with left index finger
Clockwise circle with right index finger
Anticlockwise circle with left index finger
Anticlockwise circle with right index finger
Clockwise circle with left index finger
Anticlockwise square with right index finger
Produced Bimanual Patterns
Clockwise circle with left index finger
Clockwise circle with right index finger
Anticlockwise circle with left index finger
Anticlockwise circle with right index finger
Clockwise circle with left index finger
Anticlockwise square with right index finger

Fig. 2 Comparison diagrams of expected bimanual patterns to be produced by the subjects and actually produced patterns that are distorted due to the impact of the degree of conflict between crossed and uncrossed corticospinal pathways

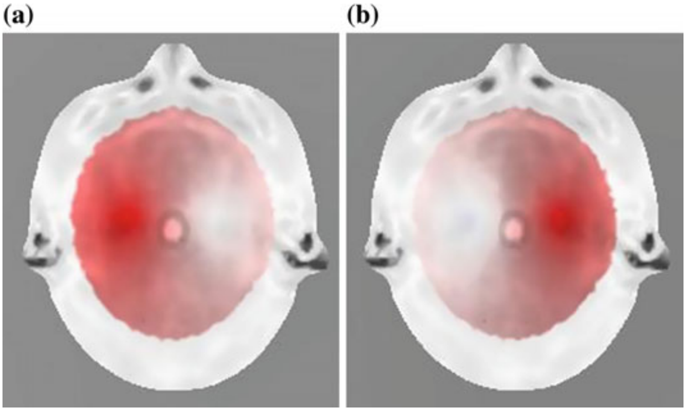


Fig. 3 Source localization technique, LORETA 2 dimensional scalp EEG maps that highlight the dominant activity distribution in (**a**) *left* and (**b**) *right* primary motor cortices

these two components depends on huge set of parameters ranging from the subject's current mental state to the accuracy of EEG setup montage. To validate and make sure that the source of the recorded EEG is primary motor cortex, a functional localisation technique, known as 'Low Resolution Electromagnetic Tomography' (LORETA) is used. By applying LORETA algorithm on the obtained EEG data, two-dimensional scalp maps are produced and displayed in Fig. 3.

From Fig. 3, it is validated that primary motor cortex is the source of the EEG at C3 and C4 10-20 positions while the subjects were executing the bimanual movements but not the volume conduction currents from the other active cortices near to the electrode surface.

Power-line noise component is removed from the EEG by using a 50 Hz notch filter. Fourth-order Butterworth band-pass filters with the following transfer function are applied to divide raw EEG data into different frequency sub bands. Filters with the frequency responses, are employed to extract Delta band (0.5–4 Hz), Theta band (4–8 Hz), Alpha band (8–13 Hz) and Beta band (13–30 Hz) of the EEG signals. Power values of all these frequency band signals of all subjects are computed for every given bimanual movement individually and these values are contrasted in the figure.

3 Results and Analysis

Right-handed persons have natural tendency to draw circles towards right side, i.e. degree of comfort is slightly high while drawing same patterns in clockwise direction compared to anticlockwise direction. This behaviour can be verified from the median band power values plots in Fig. 4. Except in delta band, in all other EEG bands average band power values in case of drawing a circle in clockwise

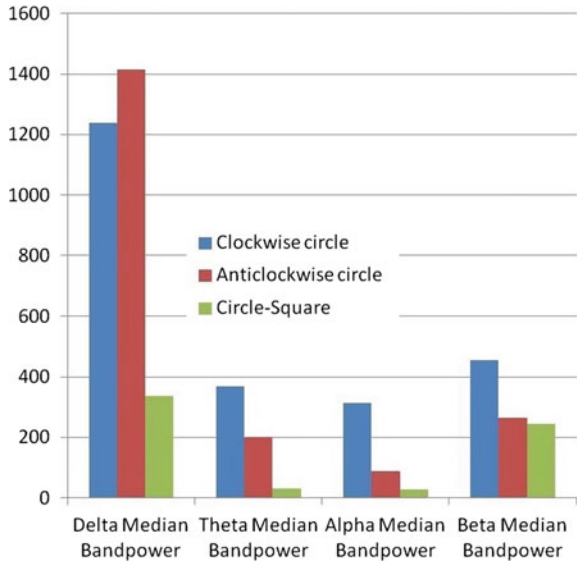


Fig. 4 Median band power values of delta, theta, alpha and beta frequency bands of primary motor cortex EEG for clockwise and anticlockwise asymmetric circle and symmetric circle-square movements

direction are higher than that of anticlockwise direction for all the right-handed subjects in this study. This inference can be interpreted as the primary motor cortex's non-delta EEG band activity is enhanced because of no-conflict arousal condition between crossed and uncrossed corticospinal pathways even if the circle drawing rate goes high. This version of interpretation reveals the direct relation between the band power of noninvasive cortical EEG and the degree of the tendency with which a movement is performed by the subjects.

Average band powers of all EEG bands, including Delta band, in case of drawing a circle with left index finger and a square with the right index finger simultaneously are substantially low compared to that of remaining two cases. Reduced all EEG band power values for this task is consistent with the fact that the low pattern drawing rate at which this task was performed by all the participated subjects. This inference represents the extent to which the severity of the conflict between crossed and uncrossed corticospinal pathways prevents the information flow to the actuator blocks.

4 Conclusion

Performing asymmetric anticlockwise bimanual movements by the right-handed subjects has resulted in considerably low median powers of non-delta EEG bands of the primary motor cortex compared asymmetric clockwise movements. Reasons for the behaviour of delta EEG median band powers in this case are yet to be exhumed. Similarly significantly low band power values of all the EEG frequency bands in circle-square drawing task are noted. All these observations insinuate that bimanual interference phenomena is not just about the information leaks in the communication lines of Carpus callosum and corticospinal pathways, but due to the feedback from these lines to the primary motor cortex, i.e. execution block of planned bimanual movements. Results of this study tell to give a serious thought to trace back the information flow in the feedback connections from the corticospinal pathways through corpus callosum to the primary motor cortex to unravel the underlying functioning principles of spatiotemporal coupling of bimanual interference movements. Further studies in this direction would be centered on studying the tuning of EEG band powers of posterior parietal cortex where planning of movements takes place, because of diverse set of continuous bimanual movements.

References

1. Mechsner, F., Kerzel, D., Knoblich, G., & Prinz, W., Perceptual basis of bimanual coordination, Nature, 414, 69–73, 2001.
2. Preilowski, B. F. Possible contribution of the anterior forebrain commissures to bilateral motor coordination, Neuropsychologia 10, 267–277 (1972).
3. Spijkers, W., Heuer, H., Kleinsorge, T., & van der Loo, H., Preparation of bimanual movements with same and different amplitudes: Specification interference as revealed by reaction time, Acta Psychologica, 96, 207–227, 1997.
4. E. Otte and H. I. van Mier, Bimanual interference in children performing a dual motor task, Hum. Mov. Sci., vol. 25, no. 45, pp. 678–693, 2006.
5. Franz, E.A., Zelaznik, H.N., & McCabe, G., Spatial topological constraints in a bimanual task, Acta Psychologica, 77, 137–151, 1991.
6. J. Gooijers, K. Caeyenberghs, H. M. Sisti, M. Geurts, M. H. Heitger, A. Leemans, and S. P. Swinnen, Diffusion tensor imaging metrics of the corpus callosum in relation to bimanual coordination: Effect of task complexity and sensory feedback, Hum. Brain Mapp., vol. 34, no. 1, pp. 241–252, 2013.
7. Carson, R. G., Neural pathways mediating bilateral interactions between the upper limbs. Brain Research, Brain Research Reviews, 49(3), 641–662, 2005.
8. Gazzaniga, M. S., Holtzman, J. D., Deck, M. D. & Lee, B. C. MRI assessment of human callosal surgery with neuropsychological correlates, Neurology 35, 1763–1766 (1985).
9. Corballis, P. M., Inati, S., Funnell, M. G., Grafton, S. T. & Gazzaniga, M. S. MRI assessment of spared fibers following callosotomy: a second look, Neurology 57, 1345–1346 (2001).
10. Filiep Debaere, Stephan P. Swinnen, Erik Be´atse, Stefan Sunaert, Paul Van Hecke, and Jacques Duysens, Brain Areas Involved in Interlimb Coordination: A Distributed Network, NeuroImage 14, 947–958 (2001)
11. U. Rokni, O. Steinberg, E. Vaadia, and H. Sompolinsky, Cortical representation of bimanual movements., J. Neurosci., vol. 23, no. 37, pp. 11577–11586, 2003.

Part III
Authentication Methods, Cryptography and Security Analysis

Probabilistically Generated Ternary Quasigroup Based Stream Cipher

Deepthi Haridas, K.C. Emmanuel Sanjay Raj, Venkataraman Sarma and Santanu Chowdhury

Abstract Presently the crypto-research based on n-quasigroup for n = 3 or higher is at its nascent stages. The recent ternary quasigroup cipher was illustrated using ternary quasigroups of order 4. Practically the ternary quasigroup needs to be of order 256. Stream ciphers to be used for real-world applications need to have ternary quasigroup of order 256. The present paper is an extension of Ternary Quasigroup Stream Cipher for practical applicability. The current work introduces the concept of probabilistically generated quasigroup. The probabilistically generated ternary quasigroup of order 256 improves the cryptographic strength of the cipher.Ternary quasigroups are more desirable options over quasigroup, but they impose serious memory constraints, particularly for large orders, e.g., 256 or more. The current study dynamically generates 3-quasigroup of the order 256 without any requisite to store them. The current study employs a selection criterion to choose suitable probabilistically generated ternary quasigroup with improved cryptographic strength.

Keywords Data encryption · Quasigroups · Latin squares · Linear equations · Stream cipher

The present research motive is to find mathematical accessories to design cryptographic primitives. Several cryptographic algorithms based on quasigroups have been developed [1–4]. The goal of present paper is to upgrade the available ciphers [5–9] on ternary quasigroup and quasigroups to improved stream cipher (based on ternary quasigroup). The current paper introduces a method for construction of probabilistic ternary quasigroup (of the order 256).

The structure of our paper: Probabilistic quasigroup generators are discussed in Sect. 1. Wherein Sect. 1.1 introduces the probabilistic quasigroup generators enlisting different generators. Section 1.2 discusses the probabilistic ternary quasigroup generators. The modified cipher of present paper is given in Sect. 2. Result and discussions are listed in Sect. 3. Section 4 gives the test cases. The test cases comprises

D. Haridas (✉) · K.C. Emmanuel Sanjay Raj · V. Sarma · S. Chowdhury
Department of Space, Government of India, Advanced Data Processing Research Institute (ADRIN), Secunderabad, Telangana 500009, India
e-mail: deepthi@adrin.res.in

P.K. Sa et al. (eds.), *Progress in Intelligent Computing Techniques: Theory, Practice, and Applications*, Advances in Intelligent Systems and Computing 719, DOI 10.1007/978-981-10-3376-6_17

of the comparative study of certain examples of the plaintext ciphertext pairs using previous cipher and the present modified ternary quasigroup based stream cipher results. Conclusions for the present work are given in Sect. 5.

1 Probabilistic Quasigroup Generation of Seed Quasigroup

1.1 Induction of Probabilistic Quasigroup Generators

Different ternary quasigroups are generated for different initial seed quasigroups. In the present work, quasigroups of order 256 have been used. If there corresponds a methodology to generate quasigroup using generalized constructions depending on some initial vectors. Such that whenever the generalized construction is reseeded it corresponds to different quasigroup. The advantage of using generalized construction for quasigroup generation is that the entire quasigroup need not be constructed at every point of time. Only the concerned element be generated which are to be used for encryption or decryption.

As per the available literature the quasigroup generators are usually deterministic in nature. The current paper proposes the probabilistic quasigroup generator to be used by our modified cipher:

Definition 1 (*Probabilistic Quasigroup Generators (PQGs)*) If the Quasigroup generator gets an additional input, of an uniform random number $r \in \mathcal{Z}$. The generator is called probabilistic quasigroup generator. If the Quasigroup generator is probabilistic then there are several different quasigroups generated from the same generator depending on the random number $r \in \mathcal{Z}$.

The following constructions are probabilistic quasigroup generators (PQGs) of order 256 worked out in the current paper:

PQG1: The Probabilistic Quasigroup Generator1 proposed in the current work: $e(x_1, x_2) = [e(x_1 - 1, x_2) - x_1]\text{mod}256$, where $e(x_1, x_2) = (n_1 \times x_1 \times x_2 + n_2 \times x_2 + n_3) mod 256$ for $x_1 = 0$, and $n_1, n_2, n_3 \in \phi(256) = \phi(2^8) \Rightarrow n_1, n_2, n_3$ should be relatively prime to $2^8 \Rightarrow n_1, n_2, n_3$ are odd numbers. Mathematical induction results: $e(x_1, x_2) = \left[e(0, x_2) - \frac{x_1 \times (x_1+1)}{2}\right] mod 256$. where $e(0, y) = (n_2 \times x_2 + n_3) mod 256$. Hence construction PQG1 is a nonlinear function in x_2.

PQG2: The Probabilistic Quasigroup Generator2 proposed in the current work: $e(x_1, x_2) = [e(x_1 - 1, x_2) + x_1] mod 256$, where $e(x_1, x_2) = (n_1 \times x_1 \times x_2 + n_2 \times x_2 + n_3) mod 256$ for $x_1 = 0$, and n_1, n_2, n_3 are odd numbers. Mathematical induction results: $e(x_1, x_2) = \left[e(0, x_2) + \frac{x_1 \times (x_1+1)}{2}\right] mod 256$, where $e(0, x_2) = (n_2 \times x_2 + n_3) mod 256$. Hence PQG2 is a nonlinear function in x_1.

PQG3: The Probabilistic Quasigroup Generator3 proposed in the current work $e(x_1, x_2) = [e(x_1 - 1, x_2) + n_4] mod 256$, where $e(x_1, x_2) = (n_1 \times x_1 \times x_2 +$

$n_2 \times x_2 + n_3)mod256$ for $x_1 = 0$, and n_1, n_2, n_3, n_4 are odd numbers. Mathematical induction results: $e(x_1, x_2) = \Big[e(0, x_2) + x_1 * n_4\Big]mod256$, where $e(0, x_2) = (n_2 \times x_2 + n_3)mod256$. Hence PQG3 is a linear function in x_1.

PQG4: The Probabilistic Quasigroup Generator4 proposed in the current work $e(x_1, x_2) = [n_1 \times x_1 + (256 - x_2)mod256]mod256$, where order of quasigroup is 256 and n_1 is an odd number. $e(x_1, x_2) = \Big[n_1 \times x_1 + e(0, x_2)\Big]mod256$, where $e(0, x_2) = (256 - x_2)mod256$. Hence PQG4 is a linear function in x_1.

PQG5: The Probabilistic Quasigroup Generator5 proposed in the current work $e(x_1, x_2) = (n_1 \times x_1 + n_2 \times x_2 + n_3)mod256$ where n_1, n_2, n_3 are odd numbers. $e(x_1, x_2) = \Big[n_1 \times x_1 + e(0, x_2)\Big]mod256$, where $e(0, x_2) = (n_2 \times x_2 + n_3)mod$ 256. Hence PQG5 is a linear function in x_1.

PQG6: The Probabilistic Quasigroup Generator6 proposed in the current work $e(x_1, x_2) = (\theta(x_1) + n_2 \times x_2 + n_3)mod256$, where $\theta(x_1) = n_4 \times x_1 + n_5$; $n_2, n_3, n_4, and\ n_5$ are odd numbers. Mathematical induction results: $e(x_1, x_2) = \Big[n_4 \times x_1 + e(0, x_2))\Big]mod256$, where $e(0, x_2) = (\theta(0) + n_2 \times x_2 + n_3)mod256$. Hence PQG6 is a linear function.

1.2 Reducible Ternary Quasigroup Constructed Using Seed Quasigroup

For each initial quasigroup generated by PQGs there are different ternary quasigroup operations. Those ternary quasigroup generators require two more quasigroup. Let the other two quasigroups be $[\gamma_1]$, $[\gamma_2]$, let it be the left and right parastrophe corresponding to $[\gamma]$. Hence whenever the Probabilistic Quasigroup Generator (PQGs) is invoked with a random initial seed, it results in a quasigroup.

$$\begin{aligned} \gamma(\gamma_1(x_1, x_2), x_2) = x_1 = \gamma_1(\gamma(x_1, x_2), x_2) \\ \gamma(x_1, \gamma_2(x_1, x_2)) = x_2 = \gamma_2(x_1, \gamma(x_1, x_2)) \end{aligned} \tag{1}$$

The resultant seed quasigroup when used in combination with other probabilistically generated quasigroups satisfying Eq. 2 generates probabilistic ternary quasigroup (PTQs). Probabilistic ternary quasigroup generators (PTQGs) are in line with PQGs in definition.

Definition 2 (*Probabilistic ternary quasigroup generators (PTQGs)*) If the ternary quasigroup generator gets an additional input, of an uniform random number $r \in \mathcal{Z}$. The ternary quasigroup generator is called probabilistic quasigroup generator. If the ternary quasigroup generator is probabilistic, there are several different ternary quasigroups generated from the same generator depending on the random number $r \in \mathcal{Z}$.

$$\psi_{TQ}(x_1, x_2, x_3) = \sigma_Q(x_1, x_2, r), \tag{2}$$

where $\psi_{TQ}(x_1, x_2, x_3)$ is a probabilistic ternary quasigroup generator, $\sigma_Q(x_1, x_2, n)$ is a probabilistic quasigroup generator and $r \in \mathcal{Z}$ is an integer.

The current paper uses the following probabilistic ternary quasigroup generator (PTQG) for construction of ternary quasigroups:

$$\begin{aligned} \alpha = \gamma(\gamma(x_1, x_2), x_3), \quad \alpha_1 = \gamma_1(\gamma_1(x_1, x_3), x_2) \\ \alpha_2 = \gamma_2(x_1, \gamma_1(x_2, x_3)), \quad \alpha_3 = \gamma_2(\gamma(x_1, x_2), x_3) \end{aligned} \tag{3}$$

2 Modified Ternary Quasigroup Based Stream Cipher (MTQSC)

The current paper's MTQS Cipher:
Setup: Let $\mathcal{M}$ = message space = set of all nonempty sequence of elements of $Q = \{m | m = m_1 m_2 \dots m_n \forall m_i \in Q\}$. i.e., if $m \in \mathcal{M} \Rightarrow m = m_1 m_2 \dots m_n$ = plaintext, $\mathcal{C}$ = Ciphertext Space $\Rightarrow c \in \mathcal{C} \Rightarrow c = c_1 c_2 \dots c_n$ = ciphertext and $\mathcal{K} = keyspace = K_1 \times K_2$, where K_1 key used to generate the initial quasigroup (of order 256) and $K_2 = b_1 b_2 \dots b_{64} b_{65}$, where b_i's are character $\forall i \in [1, 65]$. The present study does not use K_1, since the current work uses its own quasigroup (of the order 256) construction.

Let us start with an initial quasigroup or seed quasigroup, constructed as per method described in Sect. 1. Based on the initial quasigroup generated $\langle Q, \gamma \rangle \equiv \langle Q, \gamma, \gamma_1, \gamma_2 \rangle$ derive 3-quasigroup $\langle Q, \alpha, \alpha_1, \alpha_2, \alpha_3 \rangle$ which is used as seed for TQSC.

For each $j \in Q$, let f_j denote the permutation of Q defined as: $f_j = f_{b_j}, \forall j \in [1, 4]$, where $f_j = f_{b_j} = \alpha_j(b_j, b_{j+1}, x)$, $b_j \in K_2$, $1 \leq x \leq |Q|$ and $j + 1 = j + 1 mod 4$, if $(j + 1) > 4$.

Therefore (f_4, f_1, f_2, f_3) is an isotopy of 3-quasigroup $\langle Q, \alpha, \alpha_1, \alpha_2, \alpha_3 \rangle$. Hence by applying the isotopy on the seed 3-quasigroup the new isotopic 3-quasigroup $\langle Q, \beta, \beta_1, \beta_2, \beta_3 \rangle$ is defined as:

$$\begin{aligned} \beta(x_1, x_2, x_3) &= f_4(\alpha(f_1^{-1}(x_1), f_2^{-1}(x_2), f_3^{-1}(x_3))) \\ \beta_1(x_1, x_2, x_3) &= f_1(\alpha_1(f_4^{-1}(x_1), f_2^{-1}(x_2), f_3^{-1}(x_3))) \\ \beta_2(x_1, x_2, x_3) &= f_2(\alpha_2(f_1^{-1}(x_1), f_4^{-1}(x_2), f_3^{-1}(x_3))) \\ \beta_3(x_1, x_2, x_3) &= f_3(\alpha_3(f_1^{-1}(x_1), f_2^{-1}(x_2), f_4^{-1}(x_3))) \end{aligned} \tag{4}$$

Encryption: The encryption function $\mathcal{E}_K(m) = c_1 c_2 \dots c_n$ is as follows:

$$\begin{array}{lll}
b_{65} \bmod 3 = 1 & b_{65} \bmod 3 = 2 & b_{65} \bmod 3 = 0 \\
c_1 = \beta(m_1, b_5, b_6) & c_1 = \beta(b_5, m_1, b_6) & c_1 = \beta(b_5, b_6, m_1) \\
c_2 = \beta(m_2, b_7, b_8) & c_2 = \beta(b_7, m_2, b_8) & c_2 = \beta(b_7, b_8, m_2) \\
\vdots & \vdots & \vdots \\
c_{30} = \beta(m_{30}, b_{63}, b_{64}) & c_{30} = \beta(b_{63}, m_{30}, b_{64}) & c_{30} = \beta(b_{63}, b_{64}, m_{30}) \\
c_j = \beta(m_j, c_{j-2}, c_{j-1}); & c_j = \beta(c_{j-2}, m_j, c_{j-1}); & c_j = \beta(c_{j-2}, c_{j-1}, m_j); \\
\quad \forall j > 30 & \quad \forall j > 30 & \quad \forall j > 30
\end{array}$$

Decryption: The decryption function $\mathcal{D}_K(c) = m_1 m_2 \ldots m_n$ is as follows:

$$\begin{array}{lll}
b_{65} \bmod 3 = 1 & b_{65} \bmod 3 = 2 & b_{65} \bmod 3 = 0 \\
m_1 = \beta_1(c_1, b_5, b_6) & m_1 = \beta_2(b_5, c_1, b_6) & m_1 = \beta_3(b_5, b_6, c_1) \\
m_2 = \beta_1(c_2, b_7, b_8) & m_2 = \beta_2(b_7, c_2, b_8) & m_2 = \beta_3(b_7, b_8, c_2) \\
\vdots & \vdots & \vdots \\
m_{30} = \beta_1(c_{30}, b_{63}, b_{64}) & m_{30} = \beta_2(b_{63}, c_{30}, b_{64}) & m_{30} = \beta_3(b_{63}, b_{64}, c_{30}) \\
m_j = \beta_1(c_j, c_{j-2}, c_{j-1}); & m_j = \beta_2(c_{j-2}, c_j, c_{j-1}); & m_j = \beta_3(c_{j-2}, c_{j-1}, c_j); \\
\quad \forall j > 30 & \quad \forall j > 30 & \quad \forall j > 30
\end{array}$$

3 Results and Discussion

Present Ternary Quasigroup Stream Cipher has been implemented on Microsoft Visual C++ 6.0 running on E8400 @3.00 GHz with 2 GB RAM on Microsoft Windows XP. In order to show that present cipher does not fail with trivial plaintext combined with trivial key, the ciphertext results have been obtained corresponding to trivial plaintext operated with trivial key using different PQGs. Even though the current paper makes use of probabilistic generator still not all generators are suitable for cryptographic purpose. There exists a strong need to devise a methodology for selecting a suitable PQG for designing cryptographic primitives.

Proposition 1 (Quasigroup Selection Criteria) *Nonlinear probabilistic quasigroup generators are more suitable for generating quasigroups for crypto primitives. As linear probabilistic quasigroup generators when used on trivial plaintext with trivial key results in patterned ciphertext while nonlinear quasigroup constructions when used on trivial palintext with trivial key results in ciphertext without any pattern, suited for crypto-usage.*

Proof Let us consider the PQGs given in Sect. 1. Let us evaluate the results using the constructions of Sect. 1 in present MTQSC.

Setup: The trivial plaintext considered is a 10 KB file of repeated strings of 41 1's ("111111111111111111111111111111111111") and the trivial key considered is 9 1's ("111111111"). The trivial key is operated on trivial plaintext using

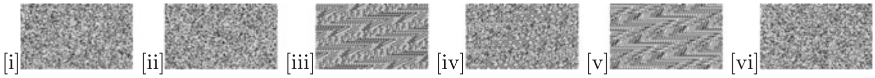


Fig. 1 Pictorial representation of ciphertext corresponding to different quasigroup constructions applied to trivial data with a trivial key: **i** Resultant Ciphertext using (nonlinear) PQG1, **ii** Resultant Ciphertext using (nonlinear) PQG2, **iii** Resultant Ciphertext using third (linear) PQG3, **iv** Resultant Ciphertext using (linear) PQG4, **v** Resultant Ciphertext using (linear) PQG5, **vi** Resultant Ciphertext using generalized (linear) PQG6 of Sect. 1

the present MTQSC with different quasigroup constructions of order 256. The different ciphertext are stored and displayed as RAW images, in Fig. 1.

PQG1: The nonlinear PQG1 of Sect. 1 when used in present MTQSC, with its variable parameters fixed as $n_1 = 25, n_2 = 19, n_3 = 23$ results in random ciphertext shown in Fig. 1i.

PQG2: The nonlinear PQG2 of Sect. 1 when used in present MTQSC, with its variable parameters fixed as $n_1 = 15, n_2 = 13, n_3 = 13$ results in random ciphertext shown in Fig. 1ii.

PQG3: The linear PQG3 of Sect. 1 when used in present MTQSC, with its variable parameters $n_1 = 21$, $n_2 = 255$, $n_3 = 1$, and $n_4 = 25$ gives repetitive patterned ciphertext as shown in Fig. 1iii.

PQG4: The linear PQG4 of Sect. 1 when used in present MTQSC, with its variable parameters $n_1 = 171$ gives patterned ciphertext as shown in Fig. 1iv.

PQG5: The linear PQG5 of Sect. 1 when used in present MTQSC, with its variable parameters $n_1 = 17$, $n_2 = 15$ and $n_3 = 13$ gives repetitive patterned ciphertext as shown in Fig. 1v.

PQG6: The linear PQG6 of Sect. 1 when used in present MTQSC, with its variable parameters $n_2 = 5$, $n_3 = 13$, $n_4 = 25$ and $n_5 = 23$ gives patterned ciphertext as shown in Fig. 1vi.

Whenever some pattern is there in ciphertext, it would be clearly visible in the corresponding ciphertext image. From Fig. 1, it is observed that linear PQGs result in patterned ciphertext whereas nonlinear PQGs result in ciphertext without any pattern. Hence the quasigroup constructions resulting in ciphertext without any pattern are desired for cryptographic usage. □

4 Test Cases

Let us consider $[\gamma]_{256\times256}$ to be constructed from constructions listed in Sect. 1. Constructing $[\gamma_1]_{256\times256}$ and $[\gamma_2]_{256\times256}$ from the initial quasigroup $[\gamma]_{256\times256}$ using Eq. 2. Define 3-quasigroup $\langle Q, \alpha\rangle \equiv \langle Q, \alpha, \alpha_1, \alpha_2, \alpha_3\rangle$ from Eq. 4 as seed for the cipher. The advantages of present cipher are listed as follows:

- Set 1, vector# 1: using PQG1 with parameters $n_1 = 25$, $n_2 = 19$, $n_3 = 23$. key $= 0x313233343536373831$, plaintext $= 0x010203040506070809$ ciphertext $= 0x7347311D6C250231B2$.

- Set 1, vector# 2: using PQG1 with parameters $n_1 = 25$, $n_2 = 19$, $n_3 = 23$. key $= 0x313233343536373831$, plaintext $= 0x112233445566777899$ ciphertext $= 0xD829F18F93CA7EC6E3$.
- Set 1, vector# 3: using PQG1 with parameters $n_1 = 17$, $n_2 = 29$, $n_3 = 97$. key $= 0x313233343536373831$, plaintext $= 0x010203040506070809$ ciphertext $= 0x92123E15D2B25B3974$.
- Set 1, vector# 4: using PQG1 with parameters $n_1 = 17$, $n_2 = 29$, $n_3 = 97$. key $= 0x313233343536373831$, plaintext $= 0x222222222222222222$ ciphertext $= 0x023876C6DFD01EB0A3$.
- Set 2, vector# 1: using PQG2 with parameters $n_1 = 17$, $n_2 = 29$, $n_3 = 97$. key $= 0x313233343536373831$, plaintext $= 0x010203040506070809$ ciphertext $= 0xD7C8ECFF3D78367A4C$.
- Set 2, vector# 2: using PQG2 with parameters $n_1 = 17$, $n_2 = 29$, $n_3 = 97$. key $= 0x313233343536373831$, plaintext $= 0x112233445566778899$ ciphertext $= 0xFC66E92ED32A7CD609$.
- Set 2, vector# 3: using PQG2 with parameters $n_1 = 17$, $n_2 = 29$, $n_3 = 97$. key $= 0x313233343536373831$, plaintext $= 0x101010101010101010$ ciphertext $= 0x5F06E20525BAFB71FA$.
- Set 2, vector# 4: using PQG2 with parameters $n_1 = 17$, $n_2 = 29$, $n_3 = 97$. key $= 0x313233343536373831$, plaintext $= 0x010101010101010101$ ciphertext $= 0xD7EED03DFC678143F3$.

5 Conclusions

The present paper works out the modified ternary quasigroup based cipher. The present paper proposes the probabilistic quasigroup generators which later leads to probabilistic ternary quasigroup generators. A novel selection criterion is developed for probabilistic quasigroup generators to be used for the cipher. From the selection criteria, it has been concluded that nonlinear quasigroup constructions are ideally fit for cryptographic usage.

Acknowledgements DH would like to thank Dr. Saibal K. Pal, Scientist-G (Scientific Analysis Group, DRDO, Metcalfe House Complex, INDIA) for his valuable advice which helped us for this research work.

References

1. McKay, BD., Rogoyski, E.: Latin squares of order 10. Electronic J. Comb.2. Available at: http://ejc.math.gatech.edu:8080/Journal/journalhome.html. (1995)
2. Markovski, S., Gligoroski, D., Bakeva, V.: Quasigroup String Processing: Part 1. Maced. Acad. of Sci. and Arts, Sc. Math. Tech. Scien. XX 1-2. pp. 13–28. (1999)

3. Markovski, S., Gligoroski, D., Stojčevska, B.: Secure two-way on-line communication by using quasigroup enciphering with almost public key. Novi Sad Journal of Mathematics. vol. 30. (2000)
4. Markovski,S., Kusakatov, V.: Quasigroup String Processing: Part 2. Contributions Sec. math. Tech. Sci. MANU XXI. vol. 1-2. pp. 15–32. (2000)
5. Petrescu, A.: Applications of Quasigroups in Cryptography. Proc. Inter-Eng. Univ. Petru Maior of Tg. Mures. Romania. (2007)
6. Petrescu, A.: n-Quasigroup Cryptographic Primitives: Stream Cipher. Studia Univ. Babes Bolyai. Informatica. Vol. LV. No 2. pp27–34. (2010)
7. Haridas, Deepthi., Venkatraman, Sarma., Varadan, Geeta.: Block Cipher Based on Randomly Generated Quasigroups. ICAMEM-2010. (2010)
8. Haridas, Deepthi., Venkatraman, Sarma., Varadan, Geeta.: Strengthened Iterated Hill Cipher for Encrypted Processing. IEEE PDGC. (2012)
9. Chakrabarti, Sucheta., Pal, SaibalK., Gangopadhyay, Sugata.: An Improved 3-Quasigroup based Encryption Scheme. ICT Innovations. (2012)

Dynamic Access Control in a Hierarchy with Constant Key Derivation Cost

Nishat Koti and B.R. Purushothama

Abstract While providing access control in a hierarchical access structure, a partially ordered set of security classes can be used to depict an access hierarchy. Data accessible to descendants of a particular security class should also be accessible to the users of that security class. Towards this, an access control scheme is proposed for providing dynamic hierarchical access control. In the proposed solution, the storage at the users is constant. The public key storage is equal to the size of the hierarchy. Also, deriving the decryption key of a descendant class involves constant cost at the users in the security class.

Keywords Hierarchical access control · Symmetric key · Dynamic classes · Constant key derivation cost

1 Introduction

In today's world, it is required that access control provides flexible access privileges. In a hierarchical structure, it is required that users at different levels in the hierarchy are given different access privileges. Users higher up in the hierarchy should be able to access information that is accessible to users lower in the hierarchy, in addition to information available at their level. Consider the example of a business organization. The various positions held in this organization may be that of a CEO, manager, staff member, and so on. In this situation, access control will have a hierarchical structure. The staff will be at a lower layer and should access only data related to their department. The managers, being at a layer higher than that of the staff, should be able to access more amount of data than that accessible to the staff members. The CEO can access all the data. Achieving such access control by applying classical

N. Koti (✉) · B.R. Purushothama
National Institute of Technology Goa, Farmagudi, Ponda 403401, Goa, India
e-mail: nishatkoti@gmail.com

B.R. Purushothama
e-mail: puru@nitgoa.ac.in

P.K. Sa et al. (eds.), *Progress in Intelligent Computing Techniques: Theory, Practice, and Applications*, Advances in Intelligent Systems
and Computing 719, DOI 10.1007/978-981-10-3376-6_18

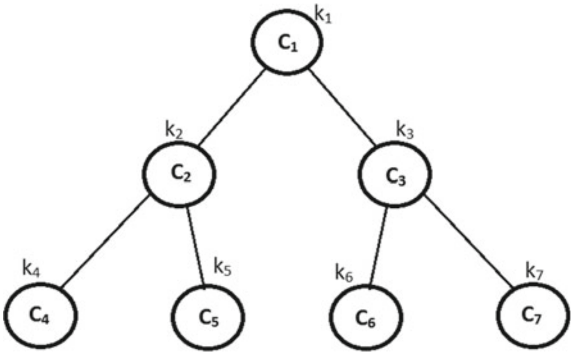


Fig. 1 Hierarchical access structure with 7 security classes

encryption would be costly to realize. Similar access control problems arise in organizations which have a hierarchical structure. These organizations may include the government, military, business organizations, and so on.

Let $C_1, C_2, \ldots C_n$ be a disjoint partition of the set of users in the system. We refer to $C_i, i = 1, 2, \ldots, n$ as a security class. Let $S = \{C_1, C_2, \ldots C_n\}$ be partially ordered by the binary relation $\leq$. The meaning of $C_i \leq C_j$ in the partially ordered set $(S, \leq)$ is that users in C_j are at a higher level in the hierarchy than users in class C_i. Hence, class C_j users have the right to access information destined to users in class C_i. However, the reverse is not allowed. When a piece of data d_k is sent to users in class C_k, users in all C_i such that $C_k \leq C_i$ can access d_k. Note that, class and security class may be used synonymously.

Consider the representation of a hierarchical structure shown in Fig. 1. As can be seen, C_1 is higher up in the access hierarchy than the other classes. Hence, users in C_1 should be able to access data destined to all its descendant classes, lower in the access hierarchy. However, information sent to C_1 should not be accessible to the users in C_2. Also, this should be achieved with minimal overhead on the users in the various classes.

Storing the secret keys of all its descendant classes by the users of a particular class, would be a trivial solution to address this problem. This would make the secret key storage dependent on the number of descendant classes. Also, while adding new classes to the access hierarchy, it should be ensured that users of the new class are not able to access information sent to its descendant classes in the past. This is called ensuring backward secrecy. Similarly, during the deletion of some existing class from the access hierarchy, the users of this deleted class should be unable to successfully decrypt any messages that will be sent to its descendant classes in the future. This is called maintaining forward secrecy. We refer to the events of adding and deleting security classes from the access hierarchy as class dynamics. Adding and removing security classes to/from the access hierarchy is a usual phenomena when considering hierarchical access structures supporting class dynamics. During these events, forward and backward secrecy needs to be maintained while creating minimal overhead on the users. Towards this, we propose a dynamic hierarchical access control scheme.

In our proposed scheme, the secret storage at the users of a class is constant and is not dependent on the number of its descendant classes. Also, the proposed solution handles the event of class addition and deletion from the access hierarchy with minimal overhead on the users of the classes. In addition to this, the cost involved in deriving a key of the descendant class is constant.

1.1 Our Contribution

We propose a solution based on symmetric key cryptography to address the problem of providing dynamic hierarchical access control. In the proposed solution, there exists a central trusted key generation center (KGC) which is responsible for managing the keys. Users in every class are required to store only their class secret keys (CSK). Users belonging to classes higher up in the hierarchy are able to access information destined to users in their descendant classes by making use of some public information. Also, the proposed solution is able to handle the event of the addition and deletion of classes from the hierarchical structure.

The following is the organization of the remainder of the paper. In Sect. 2, the existing work in this area is discussed. Section 3 covers the preliminaries required. The proposed scheme is then discussed in Sect. 4. Section 5 discusses how class dynamics are handled. The scheme is analyzed in Sect. 6 and the conclusion follows in Sect. 7.

2 Related Work

Akl et al. [1] performed the initial work in the area of access control in hierarchical structure. Some of the other publications in this area are [2–6]. The approaches in these involve a central entity which is responsible for managing the keys and other related information. The main idea in these schemes is that users in a security class are capable of deriving their descendant class' key. A solution for hierarchical access control for a distributed environment was proposed by Birget et al. [7]. In their scheme, the model consists of two hierarchies, one for the users and one for the resources. However, their scheme is not dynamic. The schemes proposed by Sun et al. [8] and Zhang et al. [9] have a high rekeying overhead during the event of a class addition or a class deletion from the access hierarchy. Also, for large hierarchies, a large number of keys are required for each class. Ferrara et al. [10] proposed an approach involving information theory for solving the problem of access control. However, their approach requires a huge number of keys to be stored by the users in each class. Also, class dynamics result in many changes. Various schemes [11–13] have been proposed based on polynomial interpolation for providing access control in hierarchies. However, large number of keys are required to be stored by the nodes in the structure. A hierarchical key management scheme was proposed by Chou et al.

[14] which was based on quadratic residues. However, this scheme does not address the problem of dynamic addition and deletion of a class from an access hierarchy. Several other schemes [15–17] have been proposed for providing hierarchical access control.

3 Preliminaries

This section describes the preliminaries required for the proposed solution for hierarchical access control.

3.1 System Model

There exists a set of users U in the system which are divided into disjoint classes, $C_1, C_2, \dots, C_n$. Let $S = \{C_1, C_2, \dots, C_n\}$. All the users in a class C_i are associated with a class secret key, CSK_i and a class encryption key, CEK_i. Also, there is a public component γ associated with every security class, which aids the users in the key derivation process. At any point in time, a security class addition or deletion from the hierarchy may take place. There exists a central trusted key generation center (KGC) which is responsible for generating the class secret keys and class encryption keys for every class. The KGC is also responsible for updating the public components in the event of a new class addition or a class deletion from the hierarchy.

3.2 Definitions

The proposed solution comprises of the following algorithms: **System Setup**, **Key Generation** and **Key Derivation**. Each of these methods is defined below.
System Setup $(\lambda) \rightarrow (\Gamma)$: The input to this algorithm is the security parameter λ and the output consists of the system parameters Γ.
Key Generation $(\Gamma, (S, \leq)) \rightarrow (\{CSK_i\}_{i \in S}, \{CEK_i\}_{i \in S}, \{\gamma_i\}_{i \in S})$: This algorithm takes the system parameters Γ and the partially ordered set $(S, \leq)$ of security classes as input, and outputs the class secret key CSK_i, the class encryption key CEK_i and the public component γ_i for every class C_i in S.
Key Derivation $(CSK_i, \gamma_j) \rightarrow (a_j)$: The inputs to this algorithm are the CSK_i of C_i and the public component γ_j of C_j. If $C_j \leq C_i$, i.e., if class C_j is a descendant of class C_i in the access hierarchy, then the output of the algorithm is the symmetric key a_j which can be used for decrypting messages sent to class C_j.

4 Proposed Solution

This section describes the proposed scheme to address the dynamic access control problem in an access hierarchy. Consider a set of users U in a system which are divided into disjoint classes $C_1, C_2, \dots, C_n$. Let a partial order $\leq$ be defined on the set $S = \{C_1, C_2, \dots, C_n\}$ such that $C_j \leq C_i$ implies that information destined to users in class C_j should be accessible to users in C_i. The following algorithms define the proposed solution.

Group Setup

The input to this algorithm is the security parameter λ. The output consists of the system parameters which comprise a cyclic group $\mathbb{Z}_p^\star$ with a prime order p. This is carried out by the KGC.

Key Generation

The input to this algorithm is Γ and $(S, \leq)$, and produces as output CSK_i, CEK_i and γ_i for every class in S. For every class $C_i \in S$, a prime $k_i \in_R \mathbb{Z}_p^\star$ and a value $a_i \in_R \mathbb{Z}_p^\star$ are selected randomly. CEK_i is set to be equal to a_i and CSK_i is set to be equal to k_i. The public component for a class C_i is generated as follows.

- A random value $\beta_i \in_R \mathbb{Z}_p^\star$ is selected.
- For all C_j such that $C_i \leq C_j$, the CSK_j's are used to compute γ_i as $\gamma_i = (\beta_i \times CSK_i \times \prod_{C_i \leq C_j} CSK_j) - CEK_i$.

Here, the value of p selected during the setup phase should be larger than the values of γ_i where $i = 1, 2, \dots, n$. Also, the value of CEK_i should be less than the values of CSK_j. The CSK_i is securely delivered to the users of class C_i. CEK_i is available in the public component γ_i, which the users can obtain using the key derivation procedure.

Key Derivation

The input to this algorithm is the CSK_i of a class C_i and the public component γ_j of C_j where C_j is a descendant class of C_i. C_i derives the CEK_j by computing $CEK_j = (CSK_i - \gamma_j) mod\ CSK_i = a_j$. Once users in C_i obtain the CEK_j of C_j, they can decrypt the encrypted information sent to C_j.

Correctness of Key Derivation

The following shows the correctness of the key derivation procedure.

$$\begin{aligned} CEK_j &= CSK_i - \gamma_j\ mod\ CSK_i \\ &= k_i - ((\beta_j \times \prod_{C_j \leq C_i} CSK_i) - CEK_j)\ mod\ CSK_i \\ &= k_i - ((\beta_j \times \prod_{C_i \leq C_j} k_j) - a_j)\ mod\ k_i \end{aligned}$$

$$
\begin{aligned}
&= k_i - ((\beta_j \times \prod_{C_i \le C_j} k_j) mod\ k_i - (a_j) mod\ k_i)\ mod\ k_i \\
&= k_i - (0 - a_j)\ mod\ k_i \\
&= k_i - (-a_j\ mod\ k_i)\ mod\ k_i \\
&= k_i - (k_i - a_j)\ mod\ k_i \\
&= a_j
\end{aligned}
$$

Working of the System: The access hierarchy is defined and the *CSK*s are made available to the users of the various classes by the KGC. Consider the hierarchical structure shown in Fig. 1. An encrypted message c_4 is sent to users in class C_4, encrypted using the $CEK_4 = a_4$. Users in classes C_1 and C_2 should also be able to decrypt c_4, as C_4 is a descendant class of C_1 and C_2. In order to be able to decrypt c_4 to recover the underlying plaintext m_4, users in C_1 and C_2 use their *CSK* to obtain the class encryption key of C_4 using the key derivation procedure. Using CEK_4, users in C_1 and C_2 proceed to perform the decryption of c_4 and obtain m_4. Users in other classes, which do not have C_4 as its descendant class will not be able to obtain CEK_4, as their *CSK* is not a part of the public component γ_4, and key derivation will not result in the correct value of *CEK*.

5 Handling Class Dynamics

This section discusses how the proposed solution handles the events of a security class addition and deletion from the access hierarchy.

5.1 Security Class Addition

Consider a hierarchy into which we need to add a new security class C_i. Let its descendants be denoted by C_j. Using the key generation procedure, the *CSK*, *CEK* and γ are computed for this new class. To maintain backward secrecy, messages sent to the descendants of this class in the past should be inaccessible to the users of this new class. In order to do so, the *CEK* and public component γ of all its descendant classes has to be updated. This updation is performed by the KGC. For all classes C_j such that $C_j \le C_i$, a new value of $a'_j, \beta'_j \in_R \mathbb{Z}_p^\star$ is selected. The KGC updates the public component of the descendant classes by computing $\gamma'_j = ((\gamma_j + a_j) \times \beta'_j \times (CSK_i)) - a'_j$. The old values of a_j are discarded and replaced by the new values a'_j. As the *CSK* of the new class is included in the public component of all its descendant classes, the users of the new class will be able to access information destined to its descendant classes too. Backward secrecy is also maintained. This is because the old public component did not contain the *CSK* of the new class. Also, none of the users in the descendant classes are required to do any modifications to their secret keys.

5.2 Security Class Deletion

Consider a scenario of deleting an existing security class C_i from the hierarchical access structure. In order to maintain forward secrecy, none of the messages that will be sent to the descendants of this class in the future should be inaccessible to the users in this class. To do this, the KGC updates the *CEK* and the public component γ of all the descendants of this class. Let the descendants of C_i be denoted by C_j. The KGC selects randomly $a'_j, \beta'_j \in_R \mathbb{Z}_p^\star$. The KGC updates the public component of the descendant classes by computing $\gamma'_j = ((\gamma_j + a_j) \times \beta'_j \times (CSK_i)^{-1}) - a'_j$. The KGC discards the old value of a_j and replaces it with the new value a'_j. Since the leaving class' *CSK* is removed from the public component of the descendant classes, the users in the leaving class will not be able to successfully decrypt any message sent to the descendant classes, thus maintaining forward secrecy. Also, in the event of collusion of users from the deleted class, the recovery of the current *CEK* is not possible. This is because every time the public component is updated, it is randomized with a new value of β. Also, none of the users are required to modify any of their secrets.

6 Analysis of the Proposed Solution

In this section, we carry out the analysis of the proposed scheme with respect to the amount of private storage, public storage and the cost involved in key derivation. Key derivation of a descendant class involves performing some operations. These operations constitute the cost involved in key derivation. The analysis is performed in comparison with the existing schemes.

Table 1 gives a comparative analysis of the existing schemes with the proposed scheme. In the table, w denotes the width of the poset representing the hierarchy, $|E|$

Table 1 Comparison with existing hierarchical access control schemes

Scheme	Private storage	Public storage	Key derivation cost
Lin [18]	$\mathcal{O}(1)$	$\mathcal{O}(\|E\|)$	$\mathcal{O}(L)$
Zhong [19]	$\mathcal{O}(1)$	$\mathcal{O}(\|E\|)$	$\mathcal{O}(L)$
Chien et al. [20]	$\mathcal{O}(1)$	$\mathcal{O}(\|E\|)$	$\mathcal{O}(L)$
Chen [21]	$\mathcal{O}(1)$	$\mathcal{O}(\|E\|)$	$\mathcal{O}(L)$
Atallah et al. [22]	$\mathcal{O}(1)$	$\mathcal{O}(\|E\| + \|V\|)$	$\mathcal{O}(L)$
D'Arco et al. [23]	$\mathcal{O}(1)$	$\mathcal{O}(\|V\|)$	$\mathcal{O}(L)$
De et al. [24]	$\mathcal{O}(1)$	$\mathcal{O}(\|E\| + \|V\|)$	$\mathcal{O}(L)$
Freire et al. [25]	$\mathcal{O}(w)$	$\mathcal{O}(1)$	$\mathcal{O}(L)$
Proposed scheme	$\mathcal{O}(1)$	$\mathcal{O}(\|V\|)$	$\mathcal{O}(1)$

represents the number of edges present in the hierarchy, $|V|$ represents the number of nodes/classes comprising the access hierarchy. L represents the length between nodes C_i & C_j in the access hierarchy where $C_j \leq C_i$ and C_i wants to obtain the CEK_j of C_j.

From the table we see that, in the proposed scheme the key derivation cost is constant. This is because key derivation involves a single modular operation. However, in the existing schemes the key derivation cost is dependent on the length between the nodes C_i and C_j where $C_i \leq C_j$ and C_j is performing the key derivation procedure to obtain CEK_i. The private storage at the users in the proposed scheme is constant and involves storing only the CSK of the particular class. The public storage involves storing the public component $\gamma_i, \forall i \in S$.

7 Conclusion and Future Work

A scheme for providing access control in a dynamic access hierarchy has been proposed. The proposed solution handles the class dynamics which involves adding and deleting security classes to/from the hierarchy. The storage at the users in the proposed scheme is constant. Unlike in the existing schemes, the key derivation cost is constant and is independent of the length between the nodes involved in the key derivation process. Also, during the event of adding or deleting a security class, only the public components are required to be updated. The users need not do any modifications to their keys to incorporate the changes. As part of future work, one can work towards designing an access control scheme for hierarchical structure where class dynamics will result in the modification of a minimal number of public components.

Acknowledgements This work is supported by Ministry of Human Resource Development, Government of India.

References

1. Akl S.G., Taylor P.D.: Cryptographic solution to a problem of access control in a hierarchy. ACM Transactions on Computer Systems (TOCS), vol. 1, no. 3, pp. 239–248 (1983)
2. Chang C.C., Buehrer D.J.: Access control in a hierarchy using a one-way trap door function. Computers & Mathematics with Applications, Elsevier, vol. 26, no. 5, pp. 71–76 (1993)
3. He M., Fan P., Kaderali F., Yuan D.: Access key distribution scheme for level-based hierarchy. Proceedings of the Fourth International Conference on Parallel and Distributed Computing, Applications and Technologies (PDCAT 2003), pp. 942–945 (2003)
4. Liaw H.T., Wang S.J., Lei C.L.: A dynamic cryptographic key assignment scheme in a tree structure. Computers & Mathematics with Applications, Elsevier, vol. 25, no. 6, pp. 109–114 (1993)

5. MacKinnon S.J., Taylor P.D., Meijer H., Akl S.G.: An optimal algorithm for assigning cryptographic keys to control access in a hierarchy. IEEE Transactions on Computers, no. 9, pp. 797–802 (1985)
6. Freire E.S.V., Paterson K.G.: Provably secure key assignment schemes from factoring. Information Security and Privacy, Springer. pp. 292–309 (2011)
7. Birget J.C., Zou X., Noubir G., Ramamurthy B.: Hierarchy-based access control in distributed environments. IEEE International Conference on Communications (ICC 2001), vol. 1, pp. 229–233 (2001)
8. Sun Y., Liu K.J.: Scalable hierarchical access control in secure group communications. Twenty-third AnnualJoint Conference of the IEEE Computer and Communications Societies (INFOCOM 2004), vol. 2, pp. 1296–1306 (2004)
9. Zhang Q., Wang Y.: A centralized key management scheme for hierarchical access control. Global Telecommunications Conference (GLOBECOM'04), IEEE, vol. 4, pp. 2067–2071 (2004)
10. Ferrara A.L., Masucci B.: An information-theoretic approach to the access control problem. Theoretical Computer Science, Springer, pp. 342–354 (2003)
11. Chang C.C., Lin I.C., Tsai H.M., Wang H.H.: A key assignment scheme for controlling access in partially ordered user hierarchies. 18th International Conference on Advanced Information Networking and Applications (AINA 2004), vol. 2, pp. 376–379 (2004)
12. Das M.L., Saxena A., Gulati V.P., Phatak D.B.: Hierarchical key management scheme using polynomial interpolation. ACM SIGOPS Operating Systems Review, vol. 39, no. 1, pp. 40–47 (2005)
13. Tsai H.M.,Chang C.C.: A cryptographic implementation for dynamic access control in a user hierarchy. Computers & Security, Elsevier, vol. 14, no. 2, pp. 159–166 (1995)
14. Chou J.S., Lin C.H., Lee T.Y.: A novel hierarchical key management scheme based on quadratic residues. Parallel and Distributed Processing and Applications, Springer, pp. 858–865 (2004)
15. Chung Y.F., Lee H.H., Lai F., Chen T.S.: Access control in user hierarchy based on elliptic curve cryptosystem. Information Sciences, Elsevier, vol. 178, no. 1, pp. 230–243 (2008)
16. Chuang Y.H., Hsu C.L.: A Robust Dynamic Access Control Scheme in a User Hierarchy Based on One-Way Hash Functions. Journal of Internet Technology, vol. 15, no. 2, pp. 197–201 (2014)
17. Wu J., Wei R.: An access control scheme for partially ordered set hierarchy with provable security. Selected Areas in Cryptography, Springer, pp. 221–232 (2005)
18. Lin C.H.: Hierarchical key assignment without public-key cryptography. Computers & Security, Elsevier, vol. 20, no. 7, pp. 612–619 (2001)
19. Zhong S.: A practical key management scheme for access control in a user hierarchy. Computers & Security, Elsevier, vol. 21, no. 8, pp. 750–759 (2002)
20. Chien H.Y., Jan J.K.: New hierarchical assignment without public key cryptography. Computers & Security, Elsevier, vol. 22, no. 6, pp. 523–526 (2003)
21. Chen T.S., Huang J.Y.: A novel key management scheme for dynamic access control in a user hierarchy. Applied Mathematics and Computation, Elsevier, vol. 162, no. 1, pp. 339–351 (2005)
22. Atallah M.J., Blanton M., Fazio N., Frikken K.B.: Dynamic and efficient key management for access hierarchies. ACM Transactions on Information and System Security (TISSEC), vol. 12, no. 3, pp. 18 (2009)
23. D'Arco P., De S.A., Ferrara A.L., Masucci B.: Variations on a theme by Akl and Taylor: Security and tradeoffs. Theoretical Computer Science, Elsevier, vol. 411, no. 1, pp. 213–227 (2010)
24. De S.A., Ferrara A.L., Masucci B.: Efficient provably-secure hierarchical key assignment schemes. Theoretical Computer Science, Elsevier, vol. 412, no. 41, pp. 5684–5699 (2011)
25. Freire E.S.V., Paterson K.G., Poettering B.: Simple, efficient and strongly KI-secure hierarchical key assignment schemes. Topics in Cryptology (CT-RSA 2013), Springer, pp. 101–114 (2013)

An Efficient LWE-Based Additively Homomorphic Encryption with Shorter Public Keys

Ratnakumari Challa and VijayaKumari Gunta

Abstract Public key encryption schemes developed based on learning with error problem became popular for homomorphic encryption, and are proved as secured schemes based on the worst-case hardness of short vector problems. Homomorphic encryption allows computations over the cipher text without decryption. Implementation of the scheme is not considered to be practical because of its larger public keys and larger cipher texts. Large public key and cipher texts require huge space in the cloud storages. However, there are some approaches proposed to shorten the cipher texts and public keys of LWE-based homomorphic encryption scheme. The objective of the paper is to introduce an idea to shorten the public keys and cipher texts in the storage. Also, support homomorphic addition operation on reduced cipher texts. The aim of the paper is not only to give the practical implementation of standard LWE-based homomorphic encryption operation (Addition) on the reduced cipher texts and also the performance of the proposed scheme.

Keywords Learning with errors • Short public keys • Short cipher texts • Pseudorandom generators • Prime factoring

1 Introduction

Regev [1] has proposed learning with errors problem (LWE) and its suitability for implementation cryptographic schemes in 2009. LWE problem has become popular and considered well suited for new research on public key cryptography. LWE problem has good features based on which it has lead to an explosion in research:

R. Challa (✉)
Department of Computer Science and Engineering, JNTUK, Kakinada, AP, India
e-mail: ratnamala3784@gmail.com

V. Gunta
Department of Computer Science and Engineering, JNTUH, Hyderabad, Telangana, India
e-mail: Vijayakumari.gunta@gmail.com

P.K. Sa et al. (eds.), *Progress in Intelligent Computing Techniques: Theory, Practice, and Applications*, Advances in Intelligent Systems and Computing 719, DOI 10.1007/978-981-10-3376-6_19

- cryptosystems implementation based on LWE is simple and fast
- computational assumptions of the system relies on well defined as hard as complex problems

LWE problems are believed to be hard problem [2] because well-known algorithm to solve the problem run in exponential time. Also, LWE problem is generalized form of well-known learning from parity with noise (LPN) problem which is believed to be complex NP problem in learning theory. As LWE problem is an extension of the LPN problem, it is also considered to be hard. Several homomorphic encryption using LWE problem is investigated and considered to be practical implementation to preserve privacy in cloud computing. Fully homomorphic encryption using standard LWE [4, 5], RLWE (Ring LWE) [5, 6], and other variants [3, 7, 8] of it are given theoretical implementation of the scheme. Practical implementation of fully homomorphic encryption scheme using LWE is considered to be complex due to the larger key size and cipher text size. In most of the cases, the schemes are not considered to be practical because of storage and time constraints. However, there are some ways to shorten the size of public keys and the cipher texts.

The paper presents a technique to greatly reduce the size of the cipher texts and public keys; also provides the method to support homomorphic addition operations over the shortened cipher texts. We proposed a method to shorten the public keys and cipher texts of the standard LWE homomorphic encryption scheme and investigate the homomorphic addition operations on the shorter cipher texts. The method is mainly implemented based pseudorandom generator using seed.

2 Preliminaries

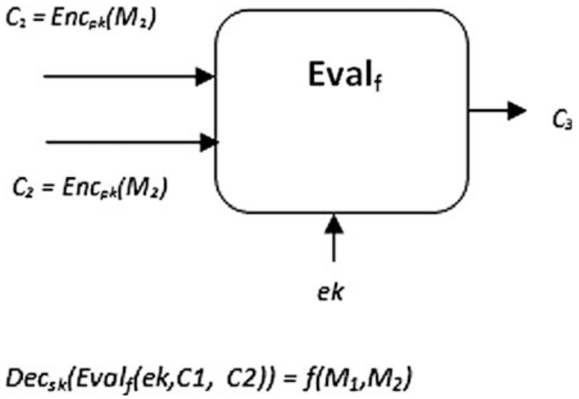
2.1 Homomorphic Encryption

Homomorphic encryption is an encryption that supports basic operations on the data which is in the encrypted form. Homomorphic encryption is very useful to perform computations on the private data which is stored on the cloud storage. Many public key cryptosystems support the homomorphic encryption [3]. Homomorphic encryption comprises four functions:

Key Generation: $\boldsymbol{(pk, sk) = KeyGen(parameters)}$
Encryption: $\boldsymbol{C = Enc_{pk}(M)}$
Decryption: $\boldsymbol{M = Dec_{sk}(C)}$
Evaluation: $\boldsymbol{C_3 = Eval_f(ek, C_1, C_2)}$

Message decrypted from the result cipher of the evaluation function is as same as the result of the function on the original message. Figure 1 shows the homomorphic encryption functions.

Fig. 1 Functions of homomorphic encryption



Homomorphic encryption supports two basic operations additions and multiplication on the encrypted data. As any function/circuit is to be implemented using the basic operations implies that any evaluation function is considered to be possible on the encrypted data without being decrypted.

Somewhat homomorphic encryption (SHE) schemes allow simple and limited number operations on the encrypted data and fully homomorphic encryption (FHE) allows both addition and multiplication over the cipher texts unlimitedly [9].

2.2 LWE-Based Homomorphic Encryption

In LWE-based homomorphic encryption scheme [4, 10], considering the parameters: dimension n and prime modulus q, LWE problem states that if a secret vector with n elements $s \in Z_q^n$ generates the vector of $n + 1$ elements and it is considered as public key $(A,\ b) \in Z_q^{n+1}$, where s is considered as secret key and $(A,\ b) \in Z_q^{n+1}$. Public key $(A,\ b)$ is computed from the secret keys as $(A, < As > + 2\ e)$.

Where the vector $A \in Z_q^n$ is an arbitrarily chosen vector with n elements and e is small random error. Based on hardness of computing secret vector s from the arbitrary number of public key samples, the security of the encryption scheme is proved. Encryption of the plain text m (0 or 1) is straight forward and simple using public key pair $(A,\ b)$, computes the cipher text C *as follows:*

$$C = (A, b) \in Z_q^{n+1},$$

where $A = A$,

$$b = b + m(mod\, q) = <As> + 2e + m(mod\, q) \in Z_q^{n+1}$$

Decryption process involves two steps to decipher the message bit *m*: First is eliminating the mask $< As >$: product of secret key *s* and coefficient vector *A* (which is part of cipher text). In the second step, an even mask is eliminated using *mod 2* operation. The decryption equation is given as follows:

$$m = (b - <AS>)mod\,2$$

Since

$$b\,\text{is} <AS> +2e + m(mod\,q)$$

In the first step, it produces 2e + *m (mod q)* and eliminating an even mask leaves the message bit as *0 or 1.*

Homomorphic encryption allows both addition and multiplication operations over the cipher texts. Let *(A, b) and (A', b')* be two cipher texts. Equations of the operations are given as follows:

Addition:

$$f_{(A+A',b+b')}(x) = f_{A,b}(x) + f_{(A',b')}(x)$$

Multiplication:

$$f_{(A'',b'')}(x) = f_{A,b}(x).f_{(A',b')}(x)$$

3 Proposed Scheme

The set of *m* public keys pairs generated from the KeyGen algorithm must be stored in the storages in order to use the public key vector *(A, b)* for encryption of the messages. Since each public key is a vector of *n + 1* integers, total space required to store *m* public keys is ***[m.(n + 1). log$_2$(q)]*** in bits. The total storage space required is reduced using various techniques of shortening public keys [11, 12]. Key sizes are reduced to very short which support implementation of LWE scheme on constrained devices.

Galbraith [12] proposed an approach to reduce the key size using pseudorandom number generator initialized with value of seed. Vector *A* is generated using pseudorandom number generator function from a seed value, public key is computed as *(A, b)* in key generation. Instead of publishing vector *A* as part of public key, corresponding seed value is to be published along with other component of public key *b*. The number of elements in the shorter public key is two: one is seed value and other is *b*. The aim of the paper is to extend Galbraith proposal of short public keys and cipher texts to support homomorphic operations. Homomorphic addition is possible on short cipher texts with a minor modification of using prime number as a seed value for pseudorandom number generator to generate the entries of the vector *A*.

The idea here is to obtain short public keys and cipher texts for LWE is that one could publish the seed (prime), rather than publishing the entire vector A. In the decryption, any user can then generate the vector A using the pseudorandom generator function from the seed value. The following are algorithms for KeyGen, encryption, addition, and decryption operations in the proposed approach.

KeyGen algorithm:

Choose parameters modulus q and dimension n
Generate secret key vector $S \in Z_q^n$ and compute public keys as follows:

1. Choose prime value p_i as seed value for vector A_i
2. Generate entries for vector $A_i \in Z_q^n$, from a seed value p using random number generator function.
3. Compute public key $< A_i, b_i >$ as $(A_i, < A_iS > + 2\ e_i)$ choosing a small random error e_i
4. Publish $< p_i, b_i >$ instead of $< A_i, b_i >$.

Encryption algorithm:

1. Choose message m_i *(1 or 0)* to be encrypted
2. Choose any k^{th} public key Pk: $< p_k, b_k>$
3. Compute cipher text C for a message bit m_i using public key Pk

$$C = <p, b> = <p_k, b_k + m_i(mod\, q)>$$

4. Store cipher text C in the storage

Homomorphic addition algorithm:

1. Choose two cipher texts to be added $C_1 = <p_1, b_1 >$ *and* $C_2 = < p_2, b_2>$
2. Compute new sum cipher C_3 from C_1 and C_2 as $< p_3, b_3 >$ where $p_3 = p_1.p_2$ and $b3 = b_1 + b_2$
3. Store sum cipher text C in the storage

Decryption algorithm:

1. Choose cipher text C: $< p, b >$ to be decrypted
2. Compute $p_1, p_2, \ldots$ from p using prime factoring technique
3. Generate vector A_i using pseudorandom generator function from the seed value p_i for all prime factors of p
4. Compute vector $A = \sum_i Ai$ for all vector A_i is corresponding to the prime factor p_i
5. Compute message m from cipher text using secret key vector S as follows

$$m = (b - <AS>)mod\, 2$$

In the decryption process, prime factors are computed from the first part of cipher text using prime factoring technique [13]. From each prime factor, generate a vector A_i using pseudorandom number generator. Corresponding elements of all vectors A_i are added to generate sum vector A and then decrypt the message eliminating inner product mask <*AS*> and another mask using *mod 2*.

4 Implementation and Results

The implementation of the short public keys and cipher texts of LWE-based homomorphic encryption is practically possible. The size of the cipher texts and public keys are observed to be reduced from $n + 1$ integers to two integers. In terms of number of bits to represent m public keys, it is reduced from $[m.(n + 1).\ log_2(q)]$ *to* $[2\ m.\ log_2(q)]$. Generation of vector A using pseudorandom number generator from a prime seed value is practically possible and is used in two functions: KeyGen and decryption. It is explored that time required for random number generator to generate n elements to construct vector A is also feasible. The approach is constructed and supported for homomorphic addition operations and the time required for each operation is presented in the Table 1. Results presented in the

Table 1 Performance time in nanoseconds for the functions of homomorphic encryption

Parameters (n dimension and q is)	Function	Standard LWE (Approx. time in ns)	Proposed method (Approx time in ns)
n = 10 q is 10 bit	KeyGen	128115	108191
	Encrypt	4346	3434
	Decrypt	92011	115316
	Addition	14248	3019
n = 100 q is 20 bit	KeyGen	731139	672566
	Encrypt	3018	3019
	Decrypt	391933	740798
	Addition	49507	2656
n = 1000 q is 30 bit	KeyGen	9498768	8626352
	Encrypt	4830	3622
	Decrypt	1823319	3800594
	Addition	176898	3019
n = 10000 q is 40 bit	KeyGen	471289365	670432042
	Encrypt	60375	10868
	Decrypt	136747730	1130967447
	Addition	1749662	1812
n = 100000 q is 50 bit	KeyGen	2.0276E+11	1.78182E+11
	Encrypt	91166	73658
	Decrypt	82212890625	1.9982E+11
	Addition	211568417	4226

table shown that KeyGen and encryption time are approximately same for standard LWE and proposed model. Addition time is reduced and decryption time is increased in the proposed model.

5 Conclusions

We proposed a scheme for exploring compact LWE-based homomorphic encryption which is suitable for devices with small memory. Experimental results prove that prime number as seed value for pseudorandom generator is used to construct the vector and provides support to perform homomorphic addition on the short cipher texts. Overhead involved to factorize the primes and reconstruct a vector *A* from the prime (seed) in decryption should be reduced and should be made feasible.

References

1. O. Regev. On lattices, learning with errors, random linear codes, and cryptography. Journal of the ACM, 56(6):34, (2009). Preliminary version in STOC'05
2. Oded Regev. The learning with errors problem (invited survey). In IEEE Conference on Computational Complexity, pages 191–204. IEEE Computer Society. (2010)
3. Craig Gentry Shai Halevi Vinod Vaikuntanathan A Simple BGN-type Cryptosystem from LWE, Proceeding of EUROCRYPT'10, 506–522, Springer LNCS. (2010)
4. Zvika Brakerski_ Vinod Vaikuntanathany Efficient Fully Homomorphic Encryption from (Standard) LWE, FOCS, (2011)
5. Zvika Brakerski Craig Gentryy Vaikuntanathanz, (Leveled) Fully Homomorphic Encryption without Bootstrapping, proceedings of ITCS'12, Pages 309–325, ACM, (2012)
6. V. Lyubashevsky, C. Peikert and O. Regev, On Ideal Lattices and Learning with Errors over Rings, in H. Gilbert(ed.) EUROCRYPT 2010, Springer LNCS 6110 (2010)
7. Zvika Brakerski1 and Vinod Vaikuntanathan, Fully Homomorphic Encryption from Ring-LWE and Security for Key Dependent Messages,. In CRYPTO, pages 505–524, (2011)
8. Shweta Agrawal, David Mandell Freeman, and Vinod Vaikuntanathan, Functional Encryption for Inner Product Predicates from Learning with Errors, Proc. in ASIACRYPT'11, SPRINGER LNCS 7073, pp 21–40, (2011)
9. Craig Gentry. A fully Homomorphic encryption scheme. PhD thesis, Stanford University, (2009). https://crypto.stanford.edu/craig
10. Ratnakumari C, Vijaya Kumari G, Sunny B, Secure Image processing using LWE Based Homomorphic Encryption, Proceedings of IEEE ICECCT, Vol 2, 804–809, (2015)
11. R. Lindner and C. Peikert, Better key sizes (and attacks) for LWE-based encryption, in A. Kiayias (ed.), CT-RSA, Springer LNCS 6558 (2011) 319–339
12. Steven D. Galbraith. Space-efficient variants of cryptosystems based on learning with errors, (2013). https://www.math.auckland.ac.nz/~sgal018/compact-LWE.pdf
13. Shor, Peter W. Polynomial-Time Algorithms for Prime Factorization and Discrete Logarithms on a Quantum Computer, SIAM J. Comput. 26 (5): 1484–1509, (1997)

An Enhanced Remote User Authentication Scheme for Multi-server Environment Using Smartcard

Ashish Kumar and Hari Om

Abstract Authentication is required to permit authorized users to access the resources and restrict unauthorized users from accessing any legal resources. The earlier schemes were used to address the security-related issues for a single server environment. Nowadays, more than one server are providing services to the users, so authentication protocols in multi-server domain are in use for real-time applications. Authentication scheme proposed by Lee et al. is susceptible to various attacks, namely forgery and server spoofing, and also unsuccessful in providing mutual authentication appropriately. Li et al. have overcome the flaws of the Lee et al.'s scheme in their scheme. Unfortunately, their scheme is not secured against the forgery attack and replay attack. To address the weakness and enhance the security, we propose a more practical scheme for authenticating a remote user in an environment consisting of multiple servers. Here we use smart card and dynamic identity of the user to fulfil all the requirements of multi-server architecture. The server requires no password table for verifying the user credentials and moreover, password can be selected freely by the user. Furthermore, performance analysis shows that our scheme provides comparatively high performance.

Keywords Password verification • Security • Smart card • Multi-server environment

A. Kumar (✉) · H. Om
Department of Computer Science and Engineering, Indian Institute of Technology (ISM),
Dhanbad 826004, Jharkhand, India
e-mail: samrata@live.com

H. Om
e-mail: hariom4india@gmail.com

P.K. Sa et al. (eds.), *Progress in Intelligent Computing Techniques: Theory, Practice, and Applications*, Advances in Intelligent Systems
and Computing 719, DOI 10.1007/978-981-10-3376-6_20

1 Introduction

Due to advancement of computation and communication technologies, the servers are growing rapidly to provide more services and reduce their individual load. In conventional client/server architectures, only one server provides services to the remote user. But in multi-server architectures, services are being provided by more than one server. To provide security to these systems, the legitimacy of a remote user accessing the services is required to be checked. Authentication is one of the most common mechanisms for this. In traditional authentication scheme, registered users are needed to log into each server using their passwords. But it is painful for any user to keep in mind a large number of passwords. Therefore, an authentication scheme is needed in which a user is required to login only once to get the services of different servers.

Till now, in multi-server environment, there have been many authentication schemes proposed. Recently, scheme [1] mentioned that scheme [2] is susceptible to server spoofing as well as forgery attack. But we have found that scheme [1] still cannot resist forgery attack as well as replay attack. Therefore, a new scheme has been proposed that solves the weakness of scheme [1].

We organize the paper as outlined below. In Sect. 2, important related schemes in multi-server domain are presented. The description of scheme [1] is provided in Sect. 3 and it is crypt analyzed in Sect. 4. Sections 5 and 6 contain the proposed scheme and its security and performance analysis respectively. Finally, we conclude the paper in Sect. 7.

2 Related Work

In 2004, a scheme for authenticating in multi-server domain has been proposed by Juang [3] that requires no table for verifying the users for their legitimacy. In this scheme, it is claimed that there is no any serious time synchronization problem and user can freely select his/her password. But scheme [4] showed that scheme [3] is not efficient and thus proposed a more efficient scheme. Afterwards, authentication scheme proposed in [5] provides access control after keeping all the features of scheme [3]. Later, a scheme [6] is designed to provide user's anonymity using the concept of dynamic ID. However, scheme [7] pointed out that scheme presented in [6] has no proper mutual authentication. Further, it is not secured against the insider's attack, masquerade attack, server and registration centre spoofing attack, and thus discussed an improvement of the scheme [6]. Thereafter, scheme [2] showed that scheme proposed in [7] has no proper mutual authentication and cannot withstand masquerade attack along with server spoofing attack, and thus proposed an improved scheme. Later, Li et al. [1] reported that scheme [2] also does not provide mutual authentication and cannot withstand server spoofing and forgery attack. Therefore, Li et al. [1] proposed a new scheme that addresses all the limitations present in the scheme [2]. In 2011, an authentication scheme [8] has been proposed that uses two server paradigms. But scheme [9] claimed that scheme [8] is not safe against stolen smart card attack, impersonation attack as well as leak of verifier attack.

3 Review of Scheme [1]

Scheme [1] uses smart card together with dynamic ID to provide authentication in multi-server environment. There are three participants involved in this scheme, registration centre (*RC*), user (U_k), server (S_j). The registration centre is a trusted party that calculates $H(SID_j||H(Q))$ and $H(P||Q)$ using P and Q. Here, Q and P are chosen secret number and registration centre's master key respectively, and SID_j is server identity. The registration centre then sends $H(P||Q)$ and $H(SID_j||H(Q))$ to server, denoted as S_j securely. U_k freely selects Id_k and PW_k as his identity and password respectively, and then picks an arbitrary number b to compute $A_k = H(b \oplus PW_k)$. After that, U_k submits A_k together with Id_k to registration centre secretly. The registration centre computes $C_k = H(Id_k||H(Q)||A_k)$, $D_k = H(B_k||H(P||Q)) = H(H(Id_k||P)||H(P||Q))$, $E_k = B_k \oplus H(P||Q) = H(Id_k||P) \oplus H(P||Q)$ and saves them in a smartcard along with (b, $H(Q)$, $H(.)$).

Thereafter, the smartcard is inserted into a reader. Then, U_k enters Id_k and PW_k for login. The smartcard calculates $A_k = H(b \oplus PW_k)$, $H(Id_k||H(Q)||A_k)$, and compares $H(Id_k||H(Q)||A_k)$ with C_k. The smartcard ends this session, if both are unequal; else U_k is identified as a legitimate user. Then the smartcard generates a nonce N_{k1} to calculate (P_{kj}, CID_k, M_1, M_2) as: $CID_k = A_k \oplus H(D_k||SID_j||N_{k1})$, $P_{kj} = E_k \oplus H(H(SID_j||H(Q))||N_{k1})$, $M_1 = H(P_{kj}||CID_k||D_k||N_{k1})$, $M_2 = H(SID_j||H(Q)) \oplus N_{k1}$. U_k sends (P_{kj}, CID_k, M_1, M_2) as login request message to server S_j. After receiving (P_{kj}, CID_k, M_1, M_2), the following steps are performed by S_j:

Step 1: S_j computes $N_{k1} = M_2 \oplus H(SID_j||H(Q))$, $E_k = H(H(SID_j||H(Q))||N_{k1}) \oplus P_{kj}$, $B_k = H(P||Q) \oplus E_k$, $D_k = H(B_k||H(P||Q))$ and $A_k = H(D_k||SID_j||N_{k1}) \oplus CID_k$. Then $\mathrm{S_j}$ computes $H(P_{kj}||CID_k||D_k||N_{k1})$ and compares it with M_1. S_j rejects the request if both are unequal. Otherwise, S_j approves the request and generates a nonce N_{k2} to calculate $M_3 = H(D_k||A_k||N_{k2}||SID_j)$ and $M_4 = A_k \oplus N_{k1} \oplus N_{k2}$. Finally, S_j sends (M_3, M_4) to U_k.

Step 2: U_k computes $N_{k2} = A_k \oplus N_{k1} \oplus M_4$, $H(D_k||A_k||N_{k2}||SID_j)$ and compares $H(D_k||A_k||N_{k2}||SID_j)$ with the message M_3. U_k ends the session if both are unequal. Otherwise, S_j will be successfully authenticated by U_k. After that U_k computes $M_5 = H(D_k||A_k||N_{k1}||SID_j)$ and sends M_5 to S_j.

Step 3: S_j calculates $H(D_k||A_k||N_{k1}||SID_j)$ and compares it with M_5. U_k is authenticated successfully if both are equal. S_j and U_k then compute a session key Sk as: $Sk = H(D_k||A_k||N_{k1}||N_{k2}||SID_j)$.

4 Cryptanalysis of Scheme [1]

4.1 Forgery Attack

Scheme [1] claimed that it can withstand forgery attack and it is not possible for any adversary E to fool S_j even after extracting the stored parameters (C_k, D_k, E_k, b, $H(Q)$, $H(.)$} of U_k's smart card. E cannot compute the valid login message (P_{kj}, CID_k, M_1, M_2) without knowing the value of A_k. To get the correct value of A_k, value of PW_k is required, which is known only to valid user U_k. However, we have analyzed that E can still masquerade as a legal user without having knowledge of PW_k. The details of forgery attack on scheme [1] are given as follow:

We assume that E intercepts the previous login message (P_{kj}, CID_k, M_1, M_2) of U_k sent to server S_j. We also assume that E has stolen the smart card of U_k and extracted the stored parameters D_k, E_k, $H(.)$ and $H(Q)$ using some technique like power analysis attack [10]. Now, E can calculate $N_{k1} = H(SID_j||H(Q))\oplus M_2$ and $A_k = CID_k \oplus H(D_k||SID_j||N_{k1})$. Thereafter, E generates an arbitrary number N_{k1}' equal to the length of N_{k1} and performs the following steps:

Step 1: E calculates $CID_k^* = A_k \oplus H(D_k||SID_j||N_{k1}')$, $P_{kj}^* = E_k \oplus H(H(SID_j||H(Q))||N_{k1}')$, $M_1^* = H(P_{kj}^*||CID_k^*||D_k||N_{k1}')$ and $M_2^* = H(SID_j||H(Q))\oplus N_{k1}'$. After that, E sends login request message (P_{kj}^*, CID_k^*, M_1^*, M_2^*) to S_j.

Step 2: After receiving (P_{kj}^*, CID_k^*, M_1^*, M_2^*), S_j computes $N_{k1}' = M_2^* \oplus H(SID_j||H(Q))$, $E_k = H(H(SID_j||H(Q))||N_{k1}')\oplus P_{kj}^*$, $B_k = H(P||Q)\oplus E_k$, $D_k = H(B_k||H(P||Q))$ and $A_k = H(D_k||SID_j||N_{k1}')\oplus CID_k^*$. S_j accepts the message if computed value $H(P_{kj}^*||CID_k^*||D_k||N_{k1}')$ is equal to M_1^*. Then S_j generates a nonce N_{k2}' and computes $M_3' = H(D_k||A_k||N_{k2}'||SID_j)$ and $M_4' = A_k \oplus N_{k1}' \oplus N_{k2}'$, and sends ($M_3'$, M_4') to E.

Step 3: Upon receiving (M_3', M_4') from S_j, E computes N_{k2}' and checks if computed value of $H(D_k||A_k||N_{k2}'||SID_j)$ is equal to M_3'. Obviously, both are equal. After that E calculates $M_5' = H(D_k||A_k||N_{k1}'|| SID_j)$ as mutual authentication message. Thereafter, E sends M_5' to S_j.

Step 4: S_j computes $H(D_k||A_k||N_{k1}'|| SID_j)$ and verifies whether they are equal to M_5' or not. Again, both are equal, so S_j successfully authenticates E. Now, both S_j and E can manage to calculate the session key as: $Sk = H(D_k||A_k||N_{k1}'||N_{k2}'||SID_j)$.

4.2 Replay Attack

If E eavesdrops the previous login message (P_{kj}, CID_k, M_1, M_2) then he/she can replay the same message to S_j. Now, S_j will verify M_1 and generate N_{k2} to calculate M_3 and M_4. After that S_j will send (M_3, M_4) to E. If somehow E manages to extract the stored parameters D_k, E_k, $H(.)$ and $H(Q)$ from U_k's smart card then U_k can easily compute $N_{k1} = H(SID_j||H(Q))\oplus M_2$ and $A_k = CID_k\oplus H(D_k||SID_j||N_{k1})$. Thereafter, E can compute $N_{k2} = A_k\oplus N_{k1}\oplus M_4$, and check whether $H(D_k||A_k||N_{k2}||SID_j)$ and M_3 are equal or not. If yes then E can calculate $M_5 = H(D_k||A_k||N_{k1}||SID_j)$ and session key $Sk = H(D_k||A_k||N_{k1}||N_{k2}||SID_j)$. Thus, scheme [1] is susceptible to replay attack.

5 Proposed Scheme

There are three entities contained in the given scheme, registration centre (RC), the user (U_k), and the server (S_j). The registration centre is a trusted party that calculates $H(P||Q)$ using Q and P. Here, Q and P are chosen secret number and master secret key of registration centre, respectively. The registration centre transmits $H(P||Q)$ to S_j securely. Our scheme consists four phases which are given as follows:

5.1 Registration Phase

For registering with the server, the user U_k and registration centre perform the steps as follows:

Step 1: U_k freely chooses Id_k and PW_k as his identity and password, respectively, and picks an arbitrary number z to calculate $X_k = H(Id_k||PW_k||z)$. Then U_k submits X_k and Id_k to registration centre securely.

Step 2: Registration centre computes $A_k = H(P||Id_k)$, $B_k = H(A_k)$, $C_k = B_k\oplus H(X_k||PW_k)$, $D_k = H(B_k||X_k)\oplus H(H(P||Q))$ and $E_k = A_k\oplus H(P||Q)$.

Step 3: Registration centre sends a smartcard to U_k which contains (C_k, D_k, E_k, $H(.)$, X_k).

Step 4: U_k enters z into his smartcard. The smartcard contains (C_k, D_k, E_k, z, $H(.)$, X_k).

5.2 Login Phase

For log into the server, the steps performed as follows:

Step 1: User inserts his smartcard into reader and inputs Id_k with PW_k. Then smartcard calculates $H(Id_k||PW_k||z)$ and compares it with X_k. If equal, the user U_k is identified as a legal user; otherwise, the session is terminated.

Step 2: Smartcard computes $B_k = C_k \oplus H(X_k || PW_k)$, $H(H(P||Q)) = H(B_k || X_k) \oplus D_k$. Then, the smartcard uses a random number N_{k1} to calculate $M_1 = B_k \oplus N_{k1} \oplus H(H(P||Q))$ and $M_2 = H(SID_j || H(H(P||Q)) || B_k || N_{k1})$.

Step 3: Smartcard transmits the message (E_k, M_1, M_2) to S_j.

5.3 Authentication Phase

In order to authenticate each other, U_k and S_j perform the steps as follows:

Step 1: S_j computes $A_k = H(P||Q) \oplus E_k$ and $N_{k1} = M_1 \oplus H(A_k) \oplus H(H(P||Q))$.

Step 2: S_j computes $H(SID_j || H(H(P||Q)) || H(A_k) || N_{k1})$ and compares it with M_2. S_j ends this session, if both are unequal; otherwise, login request is accepted.

Step 3: S_j uses a random number N_{k2} to calculate $M_3 = N_{k2} \oplus H(SID_j || N_{k1})$ and $M_4 = H(SID_j || N_{k1} || N_{k2})$. S_j sends (M_3, M_4) to user U_k.

Step 4: After receiving (M_3, M_4) from server S_j, U_k computes $N_{k2} = M_3 \oplus H(SID_j || N_{k1})$ and checks whether M_4 and $H(SID_j || N_{k1} || N_{k2})$ are equal or not. If M_4 and $H(SID_j || N_{k1} || N_{k2})$ are unequal, then U_k ends this session; else, S_j is authenticated by U_k. Thereafter, U_k sends M_5 to S_j, where $M_5 = H(SID_j || H(H(P||Q)) || N_{k2})$.

Step 5: S_j verifies whether $H(SID_j || H(H(P||Q)) || N_{k2})$ and M_5 are equal or not. If both are unequal then S_j terminates the session. Otherwise, S_j successfully authenticates U_k.

Step 6: Both S_j and U_k can manage to calculate a session key Sk as $Sk = H(SID_j || H(N_{k2} || N_{k1}))$.

5.4 Password Change Phase

In our scheme, a user can change his password without any intervention of registration centre. Following steps are performed to change the password:

Step 1: Smartcard is inserted into reader and then U_k inputs Id_k and PW_k.

Step 2: The smartcard calculates $H(Id_k || PW_k || z)$ and compares it with X_k. If both are equal, the smartcard calculates $B_k = C_k \oplus H(X_k || PW_k)$, $H(H(P|| Q)) = H(B_k || X_k) \oplus D_k$.

Step 3: U_k enters new password PW_k^{new} and a new secret number z^{new}.

Step 4: Smartcard calculates $X_k^{\text{new}} = H(Id_k || PW_k^{\text{new}} || z^{\text{new}})$, $C_k^{\text{new}} = B_k \oplus H(X_k^{\text{new}} || PW_k^{\text{new}})$, $D_k^{\text{new}} = H(B_k || X_k^{\text{new}}) \oplus H(H(P||Q))$. Then, the parameters of smartcard (C_k, D_k, X_k and z) are replaced with (C_k^{new}, D_k^{new}, X_k^{new} and z^{new}) to complete this phase. Finally the smartcard contains (C_k^{new}, D_k^{new}, E_k, z^{new}, $H(.)$, X_k^{new}).

6 Security and Performance Comparison

Security of the presented scheme together with its performance is described in this section. Its performance comparison shows that it has better performance than the existing schemes [1, 2, 7] and can resist the following attacks:

6.1 Forgery Attack

Suppose the adversary E wishes to impersonate as a valid user U_k, then he needs to create M_1 and M_2 to fool S_j. In order to compute M_1 and M_2, the adversary E must know B_k, and $H(H(P||Q))$. B_k can only be calculated as: $B_k = C_k \oplus H(X_k||PW_k)$ or $B_k = H(H(P||Id_k))$. But we have assumed that P is the secret key that is known to registration centre only, and password PW_k is secretly chosen and kept only by the user U_k. Thus, adversary is not able to impersonate as a legal user.

If E is registered himself as a valid user then he can get all stored parameters (C_E, D_E, E_E, z_E, X_E) from his/her card [10]. Now, E can calculate $H(H(P||Q))$ as: $H(H(P||Q)) = H(B_E||X_E) \oplus D_E$, where $B_E = C_E \oplus H(X_E||PW_E)$. Still, E cannot compute M_1 and M_2 of user U_k without knowing B_k.

Also, if E somehow manages to extract the stored parameters of U_k's smartcard, still E cannot calculate B_k without obtaining the value of PW_k or P. Thus, our scheme can withstand the forgery attack.

6.2 Replay Attack

If E eavesdrops U_k's login message (E_k, M_1, M_2) of previous session and replay the same message to server S_j. Then, S_j will verify M_2 and sends (M_3, M_4) to E. Even after getting the values of (E_k, M_1, M_2, M_3, M_4), E cannot calculate M_5 without knowing N_{k1}. Thus, our scheme is insusceptible to the replay attack.

6.3 Stolen Smart Card Attack

Let the U_k's smartcard has been stolen by E, who has extracted the stored parameters (C_k, D_k, E_k, z_k, $H(.)$, X_k). The adversary E cannot compute Id_k and PW_k. Also, E cannot compute the values of M_1 and M_2 without knowing the value of B_k. Thus, our scheme is secured against the stolen smartcard attack.

6.4 Server Spoofing Attack

In this scheme, it is assumed that registration centre sends $H(P||Q)$ to S_j securely. Thus, only valid S_j knows the value of $H(P||Q)$. Any adversary E that is also a legal user cannot compute $H(P||Q)$ without knowing A_E and E_E. To get the value of A_E, E must know the value of P. But P is known only to registration centre.

For a valid user U_k, only legal server S_j can compute A_k as: $A_k = H(P||Q) \oplus E_k$. After that S_j can calculate $N_{k1} = M_1 \oplus H(A_k) \oplus H(H(P||Q))$, $M_3 = N_{k2} \oplus H(SID_j|| N_{k1})$, $M_4 = H(SID_j||N_{k1}||N_{k2})$ and $Sk = H(SID_j||H(N_{k2}||N_{k1}))$, and sends (M_3, M_4) to the U_k. Here, N_{K2} is the random number chosen by S_j. Hence, in the discussed scheme server spoofing is impossible.

6.5 Proper Mutual Authentication

In the given scheme, both S_j and U_k authenticate each other properly. After receiving (E_k, M_1, M_2) from U_k, the server S_j calculates A_k and N_{k1}, and then verifies whether $H(SID_j||H(H(P||Q))|| H(A_k)||N_{k1})$ is equal to M_2. If equal, the login request message is accepted by S_j; otherwise, the login request is terminated. Thereafter, S_j generates a nonce N_{k2} and calculates M_3 and M_4, and sends (M_3, M_4) to U_k. Then, U_k calculates N_{k2} and verifies whether $H(SID_j||N_{k1}||N_{k2})$ is equal to M_4 or not. If M_4 and $H(SID_j||N_{k1}||N_{k2})$ are equal, S_j will be successfully authenticated by U_k. Else, session will be terminated by U_k. After authenticating the S_j, U_k computes M_5 and send it to S_j. After that, S_j verifies whether M_5 and $H(SID_j||H(H(P||Q))||N_{k2})$ are equal or not. If both are equal then S_j authenticates U_k and mutual authentication is completed. This scheme provides proper mutual authentication.

6.6 Anonymity

In our scheme, any server S_j cannot get the identity of any user. The user identity Id_k is protected by one-way function along with the secret key of registration centre, i.e., P. Even though server S_j can compute the value of $A_k = E_k \oplus H(P||Q)$ in authentication phase, it is infeasible to compute Id_k from A_k. Thus, our scheme provides user anonymity.

6.7 Performance Comparison

Our scheme has better security features than the existing schemes [1, 2, 7]. Table 1 shows the total computation overhead of our scheme, which less than that of the

Table 1 Performance comparison

Schemes	Computation cost of login phase	Computation cost of authentication phase	Total computation cost
Scheme [1]	7 T_h	8 T_h	15 T_h
Scheme [2]	7 T_h	9 T_h	16 T_h
Scheme [7]	6 T_h	15 T_h	21 T_h
Proposed scheme	4 T_h	10 T_h	14 T_h

schemes [1, 2, 7]. Since registration phase occurs only once, we have not considered its computation cost. The cost in our login phase is less than that of all other schemes. However, it is at higher side as compared to the schemes [1, 2], but lesser than the scheme [7], as shown in Table 1. Thus, our method provides better performance. Here, we have used T_h as the notation of time taken in performing one-way hash function.

7 Conclusions

Here, we have reviewed the Li et al.'s scheme meant for multi-server environment. Despite the claim made in that scheme, it cannot provide security against the forgery, stolen smartcard, replay, and server spoofing attacks. In this paper, we have discussed an efficient scheme for multi-server environment by a smartcard and password. Our scheme provides mutual authentication and does not need any password table for verification. Further, it has a facility to a user to choose his password at his will, provides user anonymity, and has better security features than the existing schemes.

References

1. Li, X., Ma, J., Wang, W., Xiong, Y., Zhang, J.: A Novel Smart Card and Dynamic ID based Remote User Authentication Scheme for Multi-server Environments. Mathematical and Computer Modelling. vol. 58, 1, pp. 85–95. Elsevier (2013)
2. Lee, C.C., Lin, T.H., Chang, R.X.: A Secure Dynamic ID based Remote User Authentication Scheme for Multi-server Environment using Smart Cards. Expert Systems with Applications. vol. 38, 11, pp. 13863–13870. Elsevier (2011)
3. Juang, W.S.: Efficient Multi-server Password Authenticated Key Agreement using Smart Cards. IEEE Transactions on Consumer Electronics. vol. 50, 1, pp. 251–255 (2004)
4. Chang, C.C., Lee, J.S.: An Efficient and Secure Multi-server Password Authentication Scheme using Smart Cards. International Conference on Cyberworlds. pp. 417–422. IEEE (2004)

5. Chang, C.C., Kuo, J.Y.: An Efficient Multi-server Password Authenticated Key Agreement Scheme using Smart Cards with Access Control. 19th International Conference on Advanced Information Networking and Applications. pp. 257–260. IEEE (2005)
6. Liao, Y.P., Wang, S.S.: A Secure Dynamic ID based Remote User Authentication Scheme for Multi-server Environment. Computer Standards & Interfaces. vol. 31, 1, pp. 24–29. Elsevier (2009)
7. Hsiang, H.C., Shih, W.K.: Improvement of the Secure Dynamic ID based Remote User Authentication Scheme for Multi-server Environment. Computer Standards & Interfaces. vol. 31, 6, pp. 1118–1123. Elsevier (2009)
8. Sood, S.K., Sarje, A.K., Singh, K.: A Secure Dynamic Identity based Authentication Protocol for Multi-server Architecture. Journal of Network and Computer Applications. vol. 34, 2, pp. 609–618. Elsevier (2011)
9. Li, X., Xiong, Y., Ma, J., and Wang, W.: An Efficient and Security Dynamic Identity based Authentication Protocol for Multi-server Architecture using Smart Cards. Journal of Network and Computer Applications. vol. 35, 2, pp. 763–769. Elsevier (2012)
10. Messerges, T.S., Dabbish, E.A., Sloan, R.H.: Examining Smart-card Security under the Threat of Power Analysis Attacks. IEEE Transactions on Computers. vol. 51, 5, pp. 541–552 (2002)

A Proposed Bucket Based Feature Selection Technique (BBFST) for Phishing e-Mail Classification

H.S. Hota, Akhilesh Kumar Shrivas and Rahul Hota

Abstract Phishing e-mail is a common problem faced nowadays by the e-mail users, which is an attempt to acquire sensitive information like password, credit cards details, etc. by sending malicious e-mail to the users. Classification of these types of e-mail is necessary to protect the e-mail users against harmful activities. This paper proposed to develop a classification model with the help of a new feature selection technique (FST) known as bucket-based feature selection technique (BBFST) in combination of C4.5. As the name suggested, this FST removes the feature one by one from original feature space of phishing e-mail data and puts into the three buckets based upon importance of the features as relevant feature, less relevant feature and irrelevant feature, and a new feature sub set is created. Classification technique C4.5 is then applied with data of new feature subsets and compared with existing FST. Results obtained reveal that BBFST is superior to those of existing FST with 99.008% accuracy with 12 features of phishing e-mail data.

Keywords Phishing • Bucket-based feature selection technique (BBFST) • Decision tree (DT)

H.S. Hota (✉)
Bilaspur University, Bilaspur, Chhattisgarh, India
e-mail: profhota@gmail.com

A.K. Shrivas
Dr. C.V. Raman University, Kota, Chhattisgarh, India
e-mail: akhilesh.mca29@gmail.com

R. Hota
Rajiv Gandhi Proudyogiki Vishwavidyalaya, Bhopal, MP, India
e-mail: rahulhota17@gmail.com

P.K. Sa et al. (eds.), *Progress in Intelligent Computing Techniques: Theory, Practice, and Applications*, Advances in Intelligent Systems and Computing 719, DOI 10.1007/978-981-10-3376-6_21

1 Introduction

Ever-increasing growth of phishing e-mail is creating serious problems to the e-mail users. Phishing is an activity to attempt to acquire confidential information of the users by sending forgery or malicious e-mail or to ask users to provide password of online account, credit cards detail, or any other confidential information. Phishing activities may be identified based on the content of web page in the form of HTML. There are various parts of any web page like body, name, header, etc., which may contain some malicious information which will help the classifier to identify about phishing E-mail. A link of web site as phishing e-mail is sent by changing the HTML contents slightly by the phishers [1]. So there is a need of efficient classifier to separate phishing e-mail from legitimate E-mail, at the same time the number of feature to be scanned by classifier should also be reduced as much as possible. Decision tree (DT) techniques are the popular machine learning techniques to build classification model, on the other hand feature selection techniques are to be applied to remove irrelevant features from the original feature space, many rank-based FST like Gain ratio, Info Gain, etc., are existing in the literature, which identifies and removes irrelevant feature from the original feature space.

Likarish et al. [2] have used a new anti-phishing tool called Bayesian Anti-Phishing Toolbar (B-APT) and compared B-APT with Internet Explorer and FireFox. Proposed B-APT tool has shown better performance than other. Shreeram et al. [3] have proposed genetic algorithm approach to detection of phishing web pages by using rule-based system and this rule set is used to match the hyperlink. Rahmi et al. [4] have analyzed various models like Bayesian Net, AdaBoost, DT, and Random Forest using phishing data set with two different partitions as training and testing. Accuracy of model varies from partition to partition of data set. Random Forest gives highest accuracy of 93% in case of 70–30% as training–testing partitions with hybrid feature selection technique. Almomani et al. [5] have discussed various phishing techniques to classify the phishing and non-phishing data, they also compared and discussed advantages and disadvantages of various machine learning techniques for phishing e-mail detection and prediction. Akinyelu et al. [6] have suggested Random Forest DT algorithm for phishing spam e-mail classification. They have applied random forest technique on publically available phishing data set and achieved high accuracy as 99.7% accuracy with few number of features.

This paper proposed a classifier for filtering phishing e-mail data in combination of C4.5 and proposed BBFST. Experimental setup is performed with Waikato Environment for Knowledge Analysis (WEKA) open source software using 10-fold cross validation to produce more authentic results. BBFST along with C4.5 achieved 99.008% of accuracy with 12 features. Proposed algorithm is able to remove 35 features from original feature space of phishing E-mail. Results are also compared with existing FST: Gain Ratio and found to be better.

2 Proposed BBFST

Feature selection is an important task to remove irrelevant feature from high dimensionality data set. For the experiment of this research work, e-mail phishing data set is collected from http://khonji.org web site [7]. The data set contains 8266 instances with 47 features and two classes where one class represents phishing e-mail with numeric value –1 while other class represents non-phishing e-mail (Ham) with numeric value +1. All the 47 features contain numeric value related to the contents of web page. The features of the data set are indicated from 1 to 47 similar to the sequence of its appearance in the repository data site. These two classes of data are respectively 4116 instances and 4150 instances for phishing e-mail and Non phishing e-mail (Ham).

DT [8] is popular machine learning technique used for classification as one of the important data mining task and is an extended version of Iterative Dichotomizer3 (ID3) [9] which derives rules by pruning of DT. This research work utilizes C4.5 DT technique to classify the phishing e-mail data.

Experimental framework of proposed BBFST along with machine learning based DT technique is depicted in Fig. 1. The original phishing e-mail data set as stated above with 47 features is the input of the algorithm. Algorithm picks feature one by one in random manner from the original feature space to find out importance of the feature as irrelevant, less relevant and relevant feature and puts into three different buckets. We have created three buckets in which Bucket-1, Bucket-2, and

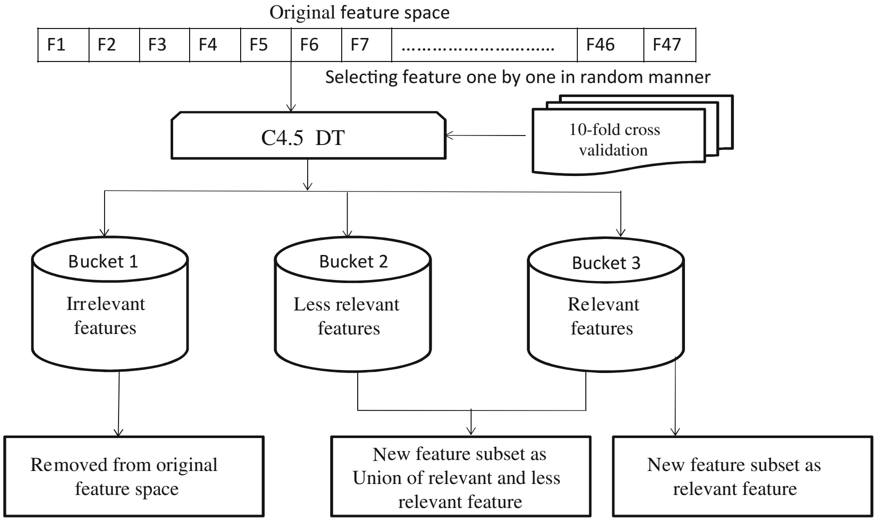

Original feature space
F1
F2
F3
F4
F5
F6
F7
…………………………
F46
F47
Selecting feature one by one in random manner
C4.5 DT
10-fold cross validation
Bucket 1
Irrelevant features
Bucket 2
Less relevant features
Bucket 3
Relevant features
Removed from original feature space
New feature subset as Union of relevant and less relevant feature
New feature subset as relevant feature

Fig. 1 Flow diagram of proposed BBFST

Bucket-3 for irrelevant features, less relevant, and relevant features, respectively. Decision is based upon the accuracy obtained using C4.5 technique, if accuracy increases after removing a particular feature from the feature space then it is kept into irrelevant feature bucket (Bucket-1), if there is no change in accuracy, it is kept into less relevant feature bucket (Bucket-2) and if accuracy decreases, feature is kept into relevant bucket (Bucket-3). In the next step irrelevant features kept into irrelevant feature bucket are completely removed from the original feature space. A new feature subset is then obtained by combining features available in relevant and less relevant feature buckets. Finally using C4.5 DT based technique results are obtained.

Experimental work is carried out using WEKA [10] machine learning tool, which consists many machine learning techniques including C4.5 DT and used in this piece of research work for classification of phishing e-mail data. In order to produce more accurate result, 10-fold cross validation technique is utilized to build classification model. In 10-fold cross validation technique, entire data set is split into 10 different folds, so that each fold will be utilized as testing set in each iteration and remaining fold will be used as training set. The result is the average of all testing and training result of all the iterations.

3 Experimental Work

Phishing e-mail classification model is built with the help of C4.5 DT technique and proposed BBFST technique using 10-fold cross validation. Initially C4.5 produces 98.88% accuracy with all 47 features of phishing e-mail data. Robustness of the classification model is checked with proposed BBFST in combination with C4.5 which segregates the features and puts it into the three buckets based on its importance. Performance of classification model is evaluated in two different cases: First with the help of new data set constructed with combination of total of 33

Table 1 Experimental results obtained in case of BBFST and Gain ratio FST using C4.5

Technique	Number of feature	Feature subset with feature ID	Accuracy
Gain Ratio—C4.5	33	{12, 3, 38, 39, 27, 20, 34, 1, 21, 13, 37, 4, 32, 22, 46, 31, 35, 15, 6, 10, 11, 41, 40, 17, 29, 45, 44, 23, 30, 42, 16, 36, 5}	98.99
	13	{12, 3, 38, 39, 27, 20, 34, 1, 21, 13, 37, 4, 32}	98.95
BBFST—C4.5	33 (Feature from Bucket-2 and Bucket-3)	{33, 19, 5, 42, 22, 4, 13, 1, 20, 27, 39, 38, 24, 18, 14, 9, 47, 8, 36, 16, 30, 23, 45, 29, 17, 40, 41, 11, 10, 35, 31, 46, 12}	99.008
	12 (Feature from Bucket-3)	{33, 19, 5, 42, 22, 4, 13, 1, 20, 27, 39, 38}	98.97

features kept in Bucket-2 and Bucket-3 while second using 12 features kept in Bucket-3, results are simulated as 99.008% and 98.97% accuracy, respectively. These results are compared with existing Gain Ratio FST combined with C4.5, simulated under similar environment and the results are found less (98.99% with 33 features while 98.95% accuracy with 12 features). However results in case of BBFST are slightly higher than that of Gain Ratio but it seems to be a competitive FST as compare to any other existing FSTs. The detail of results obtained is shown in Table 1 along with the feature ID of the phishing e-mail data set obtained in sequence from the web site [7] of the data set.

4 Conclusion

Ever-increasing number of phishing e-mail is harmful in terms of time to remove or manage as well as in terms of money also, since user feeds confidential data accidently using the web link sent by the phishers through E-mail. It is necessary to identify important HTML contents of the web link received as phishing e-mail to develop phishing e-mail classifier. Twelve out of 47 features are removed from the phishing e-mail data to build DT-based classification model using proposed BBFST. Model is built using 10-fold cross-validation technique and the results (98.97% accuracy with 12 features) obtained are slightly higher than that of Gain-Ration FST (98.95% accuracy with 12 features) simulated under similar environment.

References

1. V., Santhana Lakshmi, Vijaya, MS: Efficient prediction of phishing websites using supervised learning algorithms. International Conference on Communication Technology and System Design 2011, 30 (2012) 798–805.
2. Likarish, P., Dunbar, D. and Hansen, T. E.: B-APT: Bayesian Anti-Phishing Toolbar. IEEE Communications Society subject matter experts for publication in the ICC 2008 proceedings, (2008).
3. Shreeram, V., Suban, M., Shanthi, P. and Manjula, K.: Anti-phishing Detection of Phishing Attacks using Genetic Algorithm. IEEE, (2010) 447–450.
4. Rahmi, I., Hamid, A. and Jemal, A.: Phishing E-mail Feature Selection Approach 2011. International Joint Conference of IEEE TrustCom-11/IEEE ICESS-11/FCST-11, doi:10.1109/TrustCom.2011.126, (2011) 916–921.
5. ALmomani, A., Wan, T. C., Manasrah, A., Altaher, A., Almomani, E., Al-Saedi, K., ALnajjar, A. and Ramadass, S.: A survey of Learning Based Techniques of Phishing Email Filtering. International Journal of Digital Content Technology and its Applications (JDCTA), 6 (2012) 119–129.
6. Akinyelu, A. A., and Adewumi, A. O.: Classification of Phishing E mail Using Random Forest Machine Learning Technique. Journal of Applied Mathematics, http://dx.doi.org/10.1155/2014/425731, (2014) 1–6.

7. Web source: http://khonji.org last accessed on Feb 2016.
8. Tang, Z.H., MacLennan, J.: Data mining with SQL Server 2005. Willey Publishing, Inc., USA, (2005).
9. Pujari, A. K.: Data mining techniques, 4th edn. Universities Press (India), Private Limited, (2001).
10. Web source: http://www.cs.waikato.ac.nz/~ml/weka/ last accessed on Feb 2016.

A Novel Security Mechanism in Symmetric Cryptography Using MRGA

Bhoomika Modi and Vinitkumar Gupta

Abstract Cryptography is a primary requirement in any type of area. Cryptography is used to secure the data and communication between two parties. Some organization may have large set of data and some may have small set of data. Sometimes large data needs low security and small data needs high security. For that purpose various symmetric and asymmetric algorithms are used like DES, 3DES, AES, BLOWFISH, IDEA, RSA. These algorithms are used for encryption and decryption and measure the performance and throughput according to speed, time, and memory. In the proposed algorithm novel security mechanism is used for increased security and throughput. For security mechanism array, some arithmetic and logical operations, Magic Rectangle Generation Algorithm (MRGA) algorithm have been used in the algorithm. MRGA table is size of 16 × 24. Finally, we have done encryption and decryption using MRGA and also we have compared its throughput for different sizes of database.

Keywords Cryptography · MRGA table · Encryption · Decryption · CLS · CRS

1 Introduction

In this new generation security is the first priority. Security is increasing very much, but the problem with communication is also increasing. When user wants to secure the communication at that time he needs cryptography. For that block cipher encryption and decryption techniques are used. In symmetric cryptography only one key is used for encrypt and decrypt the data. In the cryptography plain text is

B. Modi (✉) · V. Gupta
Department of Computer Engineering, Hasmukh Goswami College
of Engineering, G.T.U, Ahmedabad, Gujarat, India
e-mail: modi.bhoomika1@gmail.com

V. Gupta
e-mail: vinit.gupta@hgce.org

P.K. Sa et al. (eds.), *Progress in Intelligent Computing Techniques: Theory, Practice, and Applications*, Advances in Intelligent Systems
and Computing 719, DOI 10.1007/978-981-10-3376-6_22

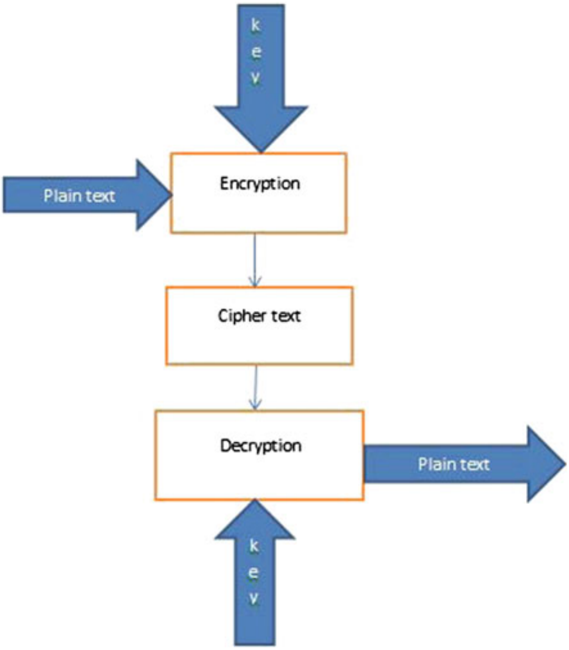


Fig. 1 Encryption–decryption process

converted in the cipher text by using key. For convert the original text into unreadable text user has to apply encryption, whereas for convert unreadable text to original text user has to apply decryption. For both encryption and decryption process in symmetric algorithm key will be same. In block cipher, symmetric techniques like DES, AES, 3DES, Blowfish are used. Some algorithm requires key generation before encryption and decryption steps. For that three main processes are required:

(1) Key generation,
(2) Encryption,
(3) Decryption (Fig. 1).

Novel security mechanism used to provide more security and high throughput. For the encryption, decryption symmetric block cipher and MRGA algorithm have been used.

2 Theoretical Background

2.1 Cryptography

Cryptography is the technique to convert the data into secret unreadable form for transferring over public network. It is used in various applications like database, communication, shopping, chatting, internet banking, and many more.

2.2 DES (Data Encryption Standard) [1, 2]

In DES algorithm there is used 64 bit of plain text and key with 56 bit is used. It also works on various shifting and XOR operation. DES has a main problem of key. Its key size is too much small so attacker can get the plain text.

2.3 3DES (Triple Data Encryption Standard) [4]

3DES is a new technique of the DES. It performs same operation as a DES but the difference is that it is encrypted three times so it is more secure than DES. The main problem is that it uses three time level of encryption which becomes very much slower than other method.

2.4 AES (Advanced Encryption Standard) [4, 5, 8]

AES is founded by Rijndael. AES has block size of 128 bits and key sizes of 128, 192, and 256 bits with 10 rounds, 12 rounds, and 14 rounds. In this algorithm XOR, mix columns, shift rows, add round key operations are performed. This algorithm suffers from brute force attack because if attacker has dictionary then he can easily break the word which is the key.

2.5 Blowfish [4]

Blowfish is the fastest and secure than above three algorithms. It has 32–448 bits of key length which is changeable, and its block size is 64 bits. Main advantage of the Blowfish is that it is openly available and not payable. These all have some weak points. Like long range of key provides high security than short. Very typical structure increases execution time.

3 Proposed Scheme

3.1 Problem Definition

In literature, author has applied symmetric encryption and decryption algorithm on the 16 characters (128 bits). The key size was also same as plain text. He has used arithmetic and logical expressions for security purpose. He has repeat some steps for more security, but that steps were not fixed so it may be problematic. So from that discussion we can say that it should be improved by using some different methods.

3.2 Proposed Method Using MRGA

We have used 16 characters of plain text and 128 bits of key for encryption and decryption processes. We have used symmetric algorithm for that process. For key generation of size 128 bits we have used Diffie–Hellman algorithm. 16 characters of plain text are converted in ASCII values and then that ASCII values are converted in MRGA values of 144 bits. MRGA is the Magic Rectangle Generation Algorithm which is size of 16 × 24. It provides polyalphabetic advantage in which for same character like in "HELLO", for both "L" it gives different values. After that encryption and decryption process takes place using arithmetic and logical operations. For odd matrix generation we have used (4 × 6) basic matrix of odd, and for even matrix we have used (4 × 6) basic matric of even. By combining these four (4 × 6) matrices there becomes (8 × 12) matrix and by combining these four (8 × 12) matrices there becomes 16 × 24 matrix which is our final table.

3.3 Algorithm for MRGA [15]

Input: maximum and minimum value
Output: singly even magic rectangle
Method

```
    $Min_{start}$ ← $MR_{start}$
    $Max_{start}$ ← $MR_{start}$-4
    i=1
    For i<=n DO
    Begin
    Call MR 4x6 fillorder ($Min_{start,}$ $Max_{start}$)
    Select the $Min_{start}$ and $Max_{start}$
    End
```

MRGA 16 × 24 is created by using these four basic tables (Table 1, 2):

```
  MR1(16X24) ← MR1(8X12)||MR3(8X12)
               ||MR2(8X12) ||MR4(8X12)
```

Table 1 Table for 4 × 6 basic matrix (odd) [15]

Max_{start}	*(+2)	*(+4)	−6	−16	*(+16)
*(+8)	−10	−12	*(+14)	*(+24)	−24
−14	*(+12)	*(+10)	−8	−30	*(+30)
*(+6)	−4	−2	*Min_{start}	*(+22)	−22

Table 2 Table for 4 × 6 basic matrix (even) [16]

Max_{start}	*(+2)	*(+4)	−6	−14	*(+14)
*(+8)	−10	−12	*(+14)	*(+6)	−6
−14	*(+12)	*(+10)	−8	−4	*(+4)
*(+6)	−4	−2	*Min_{start}	*(+12)	−12

3.4 Proposed Encryption Flowchart

See Fig. 2.

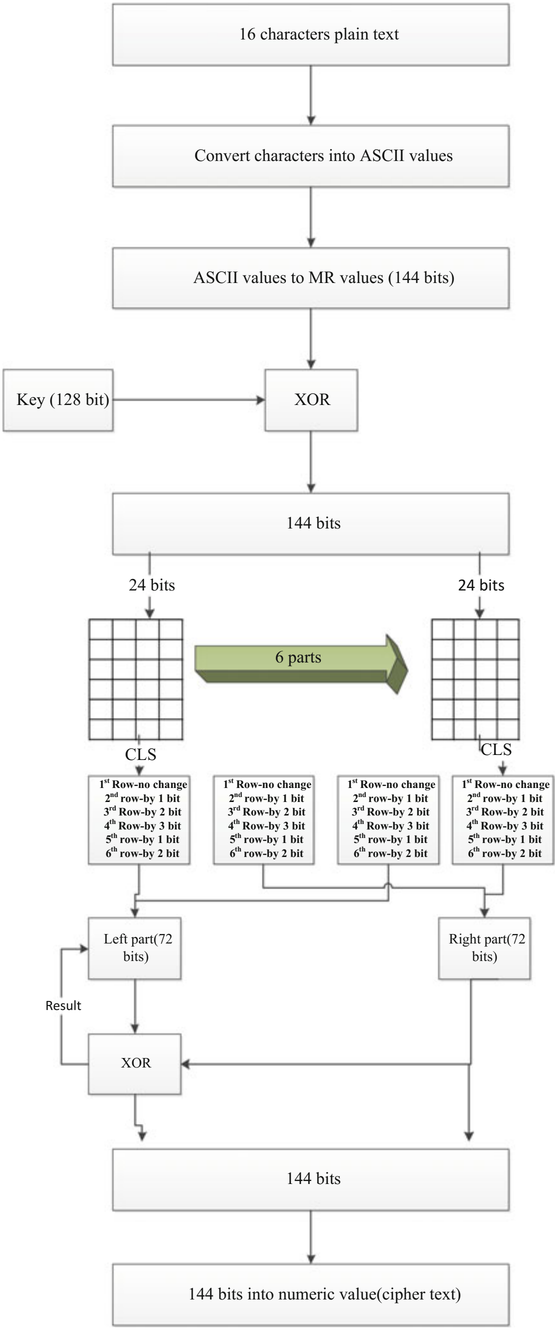


Fig. 2 Proposed encryption flowchart

3.5 Proposed Decryption Flowchart

See Fig. 3.

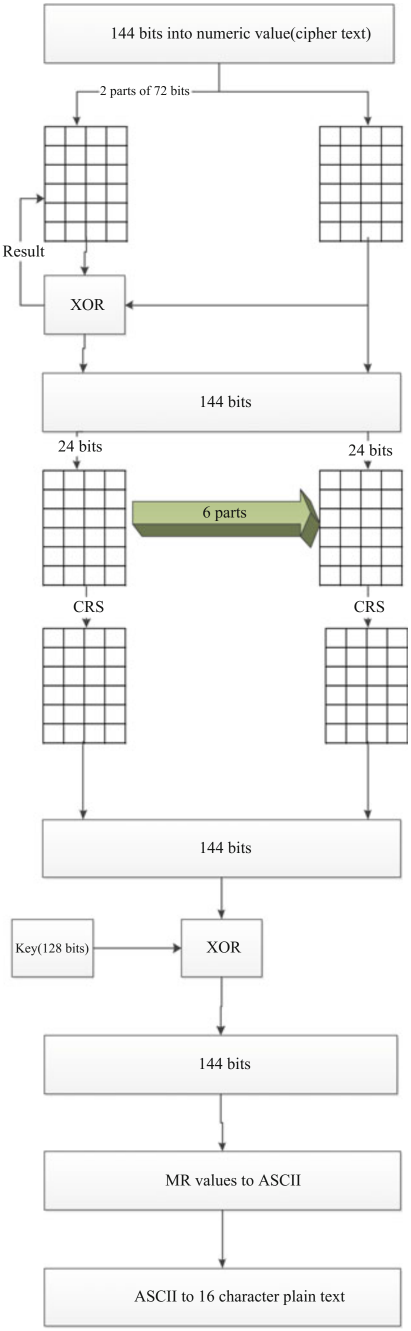


Fig. 3 Proposed decryption flowchart

4 Implementation Methodology and Results

For implementation purpose Netbeans IDE 7.0 version has been used. We have taken the 16 characters of plain text and 128 bits of key and that 16 characters are converted into ASCII values and then that ASCII values are converted into MR values from MRGA (16 × 24). For key generation we have used Diffie–Hellman algorithm. And then by applying some arithmetic and logical steps encryption and decryption processes have been done.

This is the basic algorithm and we have applied this algorithm on different sizes of database like 12 KB, 13 KB, 16 KB, and on 31 KB. We have also shown the throughput on these databases.

For this algorithm hardware requirement is CORE 2 DUO PROCESSOR with 1 GB of RAM.

Y-axis indicates time in milliseconds for each database
X-axis indicates database size (Fig. 4, Table 3).

Fig. 4 Comparison chart

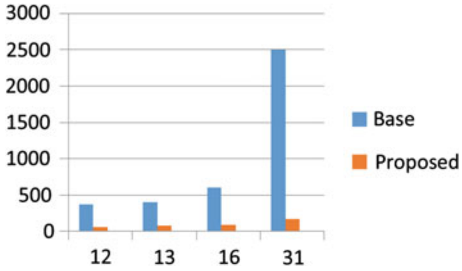


Table 3 Table of different databases with previous database

Database	Previous work (ms)	Proposed work (ms)
12	374	63
13	406	78
16	609	94
31	2496	171

5 Conclusion and Future Work

The aim of the cryptography is to transfer data in very secure manner between two parties over a network. We have proposed "The novel security mechanism in symmetric cryptography using MRGA" used to overcome some disadvantages. It gives high throughput and is more secure. It has easy process for encryption and decryption. For security purpose we have used Magic Rectangle Generation Algorithm (MRGA) of size 16 × 24. Main benefit of MRGA algorithm is that it provides more security in starting Min, Max values because those values are transfer to the receiver in encrypted form. So attacker cannot know that values even he has table. Mathematical calculation is live example for cryptography process. In future we will try to apply this algorithm on the large database and also we will compare its throughput with other work.

References

1. Nadeem, A.; Javed, M.Y. "A Performance Comparison of Data Encryption Algorithms" Information and Communication Technologies, 2005. ICICT 2005. First International Conference Publication Year: 2005, Page(s): 84–89.
2. NeetuSettia. "Cryptanalysis of modern Cryptography Algorithms" International Journal of Computer Science and Technology. December 2010.
3. S.R. Kumar, E. Pradeep, K. Naveen and R. Gunasekaran, "A Novel Approach for Enciphering Data of Smaller Bytes", International Journal of Computer Theory and Engineering, 2(4), 1793–8201, pp. 654–659, 2010.
4. J Thakur, N kumar. "DES, AES, and Blowfish: Symmetric Key Cryptography Algorithms Simulation Based Performance Analysis". International Journal of Emerging Technology and Advanced Engineering Website: http://www.ijetae.com (ISSN 2250-2459, Volume 1, Issue 2, December 2011).
5. Ramesh G and Umarani R, "A Comparative Study of Six most Common Symmetric Encryption Algorithms across Different Platforms", International Journal of Computer Applications, (Vo1. 46, No. 13, May 2012).
6. Akash Kumar Mandal and Mrs. Archana Tiwari, "Performance Evaluation of Cryptographic Algorithms: DES and AES", 2012 IEEE Students' conference on Electrical, Electronics Computer Science, 2012.
7. R. Sircar, G. Sekhon, N. Nath "Modern Encryption Standard (MES): Version-II", (978-0-7695-4958-3/13, DOI 10.1109/CSNT.2013.111, 2013, IEEE).
8. Thomas Fuhr, Eliane Jaulmes, "Fault Attacks on AES with Faulty Ciphertexts Only", (978-0-7695-5059-6/13, DOI 10.1109/FDTC.2013.18, 2013, IEEE).
9. D. Nilesh, Malti N, "The New Cryptography Algorithm with High Throughput", ICICI- 2014, jan 3–5 (978-1-4799-2352-6/14, IEEE).
10. S. Ramanujam, M. karuppiah, "designing an algorithm with high avalanche effect", IJCSNS International Journal of Computer Science and Network Security, (VOL.11 No. 1, January 2011).
11. Srinivasarao D, Sushma Rani N, "analyzing the superlative symmetric cryptographic encryption algorithm", (ASCEA), (Volume 2, No. 7, July 2011, *JGRCS*).
12. B Owaidat, Ramzi Jaber, "error correction capabilities in block ciphers", ACTEA (2012, *IEEE*).

13. R Patidar1, Rupali Bhartiya, “Modified RSA Cryptosystem Based on Offline Storage and Prime Number”, (978-1-4799-1597-2/13, 2013, *IEEE*).
14. Nedhal A. Al-Sayid, “Database Security Threats: A Survey Study”, CSIT, ISBN: 978-1-4673-5825-5 (2013, *IEEE*).
15. Dr D.I. George, J. sai, “Enhancing security level for public key ecryptosystem using MRGA”, 2014 (978-1-499-2877-4/14, *IEEE*).
16. Hardik Gandhi, “A research on enhancing public key cryptography by MRGA with RSA and N-prime”, IJIRST (volume 1, Issue 12).
17. William Stallings, “Cryptography and network security”.

Techniques for Enhancing the Security of Fuzzy Vault: A Review

Abhay Panwar, Parveen Singla and Manvjeet Kaur

Abstract Biometric Systems are the personal identification systems that use behavioral and physiological characteristics of a person. One of the main concerns in biometrics systems is template security. Fuzzy vault, a bio-cryptosystem, is used to provide security to the stored templates. Fuzzy vault has proven to be a very good security technique, nonetheless it lacks in providing revocability and security against correlation attacks. Thus for the enhancement of the security of fuzzy vault and to overcome the limitation of correlation attack, techniques like hybrid model and multimodal biometrics can be used. This paper gives a review of the above mentioned techniques, viz. hybrid and multimodal, and how they can be effective in enhancing the security of the system.

Keywords Biometric security · Fuzzy vault · Multimodal · Hybrid

1 Introduction

Biometrics is a way in which physical and behavioral attributes are used for identification of a person instead of passwords and ID cards. Identity management is one of the critical issues that are faced by most organizations. Thus the use of biometric, for such management, as a robust identification system is justified [1]. Preventing data theft by an impostor is one of the reasons to implement biometric security systems. Previously, knowledge-based and token-based methods such as passwords and ID cards, respectively, have been used as security mechanisms. But the problem with these methods is that they can be easily forgotten, lost, or stolen. Security breach has always been a problem in any authentication system and same

A. Panwar (✉) · P. Singla · M. Kaur
Computer Science and Engineering Department, PEC University of Technology,
Chandigarh 160012, India
e-mail: abhay.panwar4@gmail.com

P. Singla
e-mail: parveen7300@gmail.com

P.K. Sa et al. (eds.), *Progress in Intelligent Computing Techniques: Theory, Practice, and Applications*, Advances in Intelligent Systems and Computing 719, DOI 10.1007/978-981-10-3376-6_23

is the case with biometric authentication systems. The security attacks can occur at various stages in the system namely, *(i) sensor or image acquisition stage, (ii) feature extraction stage, (iii) matching stage, (iv) stored template stage, and (v) decision stage* [2]. The attack on the stored templates is one such attack that can be potentially damaging for any biometric system. Template is the compressed form of the enrolled biometric that only contains the required and unique features of that biometric sample. Storing the biometric data in the form of a template reduces the storage requirement to a larger extent. Some of the attacks on a template include: *(i) replacing an original template by an intruder template, (ii) creating a physical spoof from the stolen template that can be used to access any system that might use the same biometric*. There are mainly two types of schemes to secure the template, that are: *(i) feature transformation scheme and (ii) bio-cryptosystem scheme.*

1.1 Feature Transformation

Feature transformation is a scheme in which a biometric template is transformed using some transformation function. This transformed template is then stored in the database. The transformation function used can be of two types based on its characteristics, i.e., *(i) invertible (salting) and (ii) noninvertible transforms*. If a template is transformed using an invertible transformation function with the help of a key, then if the key and the transformed template are compromised then the regeneration of the original template could be possible. Therefore, this scheme is secure until the secrecy of the key is maintained [2, 3]. On the contrary if we apply a noninvertible transformation function (that is hard to invert), even if the key and the transformed template are compromised, the original template cannot be regenerated thus increasing the security of the system manifolds.

1.2 Biometric Cryptosystems

Biometric Cryptosystems can be used for various purposes. Mainly they had the use of making the cryptographic key secure by making use of the features from biometric template or generation of cryptographic key using the features from biometric template. But now, biometric cryptosystems can be put into another use of template protection. Some information from the template called "*helper data*" is publically stored in case of a biometric cryptosystem [2, 3]. The way helper data is chosen is such that it does not disclose the information of the actual template but contains the information good enough to perform a match. The extracted key is then verified for validity/invalidity for authentication. Since intra-user variations are inevitable which can be caused due to various reasons like orientation, alignment or channel noise, error correcting codes are used. Biometric cryptosystems are mainly classified into two categories, namely: *(i) key-binding cryptosystems and* *(ii) key-generation cryptosystems*. In key-binding cryptosystems a key, that does not depend on biometric features, is bound with the biometric template which results in helper data [2]. An authentication is said to be successful if the correct key is extracted from helper data during matching. On the other hand key generation biometric cryptosystem is a technique in which generation of the key happens directly from the query biometric features and helper data which in turn is itself generated from the biometric template. A better scheme, called hybrid scheme, can be formulated by making use of different approaches simultaneously, one after another. One such example would be the use of key-binding and salting together. One of the biometric cryptosystems is Fuzzy Vault Scheme described in the next section.

2 Fuzzy Vault

Fuzzy Vault lies among one of the best approaches used for key-binding biometric cryptosystem which is designed to secure biometric features. Fuzzy Vault only takes input in the form of unordered set. Biometrics that has features in ordered set must be converted into unordered set to use fuzzy vault [4, 5]. Let us suppose that B is a biometric template containing *f* feature points. A secret key K is selected and is encoded into a polynomial P with degree *d*, say, in the form of coefficients. Now, biometric template features are projected onto the polynomial. To hide the genuine feature points some random points are added that do not lie on polynomial P. These random points are called *Chaff Points*. The combined set of points forms the helper data or vault V shown in Fig. 1. Identification of genuine points from the chaff points in vault V is difficult without the presence of user, and hence the template is secure. At the time of authentication when the user provides a biometric query B′, the key K is regenerated only if B′ substantially overlaps with B and the authentication is successful. Error correcting technique is used to deal with intra-user variations. On the contrary it is infeasible to regenerate K if B and B′ do not sufficiently overlap and the authentication is unsuccessful as shown in Fig. 2 [6].

2.1 Various Fuzzy Vault Schemes

1. The authors Juels et al. [4] have proposed a Fuzzy Vault scheme for security of template. This scheme has two important features, i.e., it can take arbitrary order set and prove information-theoretic security bounds over some nonuniform distributions. In this scheme an unordered set, say, A is used to lock a secret key K. The key K is selected and is encoded into a polynomial in the form of coefficients. To hide the genuine feature points some random points are added that do not lie on polynomial. These random points are called Chaff Points. The combined set of points forms the helper data or vault. At the time of

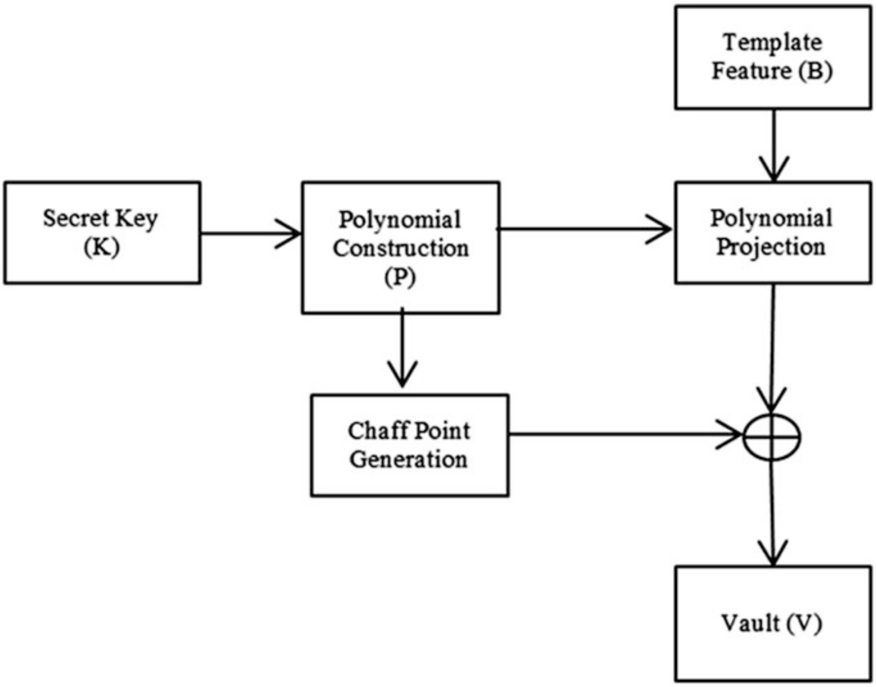


Fig. 1 Encoding phase of fuzzy vault

authentication when the user provides a biometric query B, the key K is regenerated only if B substantially overlaps with A and the authentication is successful. On the contrary it is infeasible to regenerate K if A and B do not sufficiently overlap and the authentication is unsuccessful. Reed Solomon codes are used as error correcting codes to reconstruct the polynomial. The basis of the non-vulnerability of this scheme is the infeasibility to reconstruct the polynomial.

2. Secure Smartcard-based fingerprint authentication has been proposed by the author Clancy et al. [5]. The basis of this scheme is the fuzzy vault proposed by Juels and Sudan [4]. In this paper, sets of multiple minutiae location for each finger are used. First the canonical positions of minutiae points are found which are used as input set elements for fuzzy vault. To increase the security of vault, maximum number of chaff points are added. It is assumed, in this scheme, that fingerprints are pre-aligned. Reed Solomon codes are used to reconstruct the polynomial.
3. The authors Uludag et al. [6] have proposed cryptography construct using fuzzy vault with fingerprint minutiae data. This scheme secures secret encryption key with the fingerprint data only to be used by a valid user. A 128-bit secret key can be secured with fingerprint minutiae. CRC is used for error correction and detection of polynomial and Lagrange interpolation is used to reconstruct the

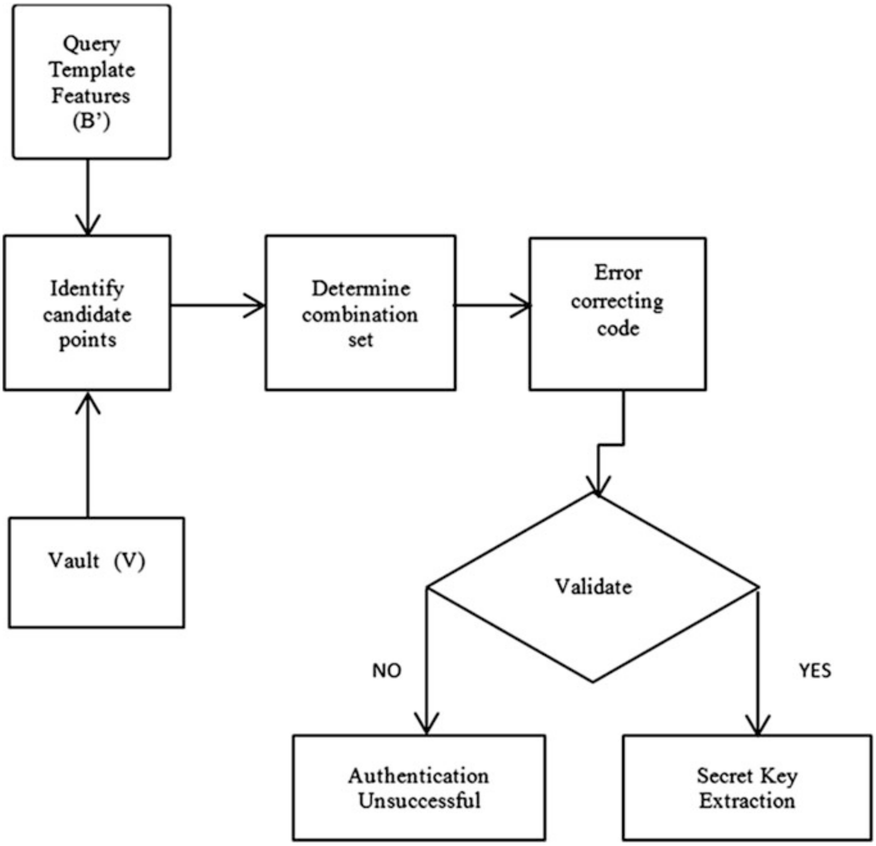

Query Template Features (B')
Identify candidate points
Determine combination set
Error correcting code
Vault (V)
Validate
NO
YES
Authentication Unsuccessful
Secret Key Extraction

Fig. 2 Decoding phase of fuzzy vault

polynomial instead of Reed Solomon codes. During encoding CRC bits are added in secret data and polynomial is constructed. Minutiae feature list is projected on polynomial and random points (chaff points) are added to securely hide the genuine points. A combined set of chaff points and genuine points construct a Vault. At the time of decoding, the vault points and the Query Minutiae List are compared. The decoding requires n + 1 unique projections to decode an n degree polynomial. All the combinations containing n + 1 points from the set of minutiae list are generated. Lagrange interpolation method is applied on each possible combination of n + 1 points to reconstruct the polynomial and that polynomial is validated by CRC error correction and detection. When applying CRC, remainder zero signifies the presence of errors while a nonzero remainder marks the error absence.

2.2 Limitations of Fuzzy Vault

In a scenario where more than one vault make use of the same biometric data, the security of vault can be jeopardized. This compromise takes place in the form of a correlation attack. This happens when an intruder correlates the values in different vaults that have been constructed from the same biometric data thus identifying the genuine points [7]. This is one of the reasons why same biometric data is denied to be used for the reconstruction of the vault, thus being non-revocable. To overcome these limitations, two methods can be used, Multimodal Approach and Hybrid Approach.

3 Multimodal Approach

In multimodal biometric systems, the biometric information is processed in the form of various factors or sources [8, 9]. For example, multiple-sensor, multiple-algorithm, multiple-instance, multiple-sample, etc. When this scheme is used with a fuzzy vault, it helps in enhancing the security of that biometric system. For example consider that a person enrolls his or her biometric traits for two different systems. One system takes into account fingerprint and face as biometric traits while the other system takes fingerprint and iris as biometric traits. So even if the intruder gets access to the vaults, only the fingerprint points can be correlated thus keeping the vault secure. Below are the various multimodal schemes:

1. The authors A. Nandakumar et al. [10] have proposed a scheme which is used to secure multiple templates of a user in a multi-biometric system in the following ways: (i) convert the different biometric data into same representation so that they are capable to be fused together, (ii) a single multi-biometric template is constructed by fusing the templates on a feature level, and (iii) apply fuzzy vault to secure the multi-biometric template. Feature set is extracted from fingerprint and iris images. As Fuzzy vault only accept unordered set so to convert the iris code from ordered to unordered, invertible transformation function is applied with the help of transformation key. Transformed iris template along with high curvature points of fingerprint is stored with the vault as helper data. Actual points in vault are hidden by adding random points (chaff points) to increase the security of vault. During authentication Lagrange interpolation is used to reconstruct the polynomial and are checked based on CRC error detection and correction scheme.
2. V.S. Meenakshi et al. [11] have designed a technique to secure the multi-biometric template. Iris, Retina and Fingerprint are used as biometrics. An invertible transform is applied using a password to the extracted minutiae points from iris, retina and fingerprint. A combination of the transformed and fused points from these modalities is provided as an input to the fuzzy vault. The

invertible transform using password on biometric features increases the security of the vault.

3. Li Yuan et al. [12], in their work, have proposed a hybrid multi-biometric template protection method with face and ear modality. In the enrollment stage, a noninvertible transformation is first applied on the real valued face and ear fused templates to convert this ordered point set to unordered point set. Then these vault points are encrypted and stored in the database together with some random points (chaff points). Addition of these random points hides the genuine points to increase the security of vault. In the authentication stage the helper data created from the enrolled biometrics features, and the query templates are matched against each other. The output of the final matching may result in: (i) authentication success and releasing the key or (ii) authentication failure if enough overlap is not found in the two templates.

4 Hybrid Approach

A hybrid approach in biometric systems is a combination of feature transformation and bio-cryptosystems. When using feature transformation, only the transformed templates are matched against each other. Due to intrauser variations the matching performance can decrease as the transformed template will not be same as the query template. Similarly when a system uses a bio-crypto system alone, it may be vulnerable to correlation attacks. Thus by combining these two approaches, i.e., into a hybrid model security and performance can both be managed. Below are discussed the various hybrid approaches:

1. A hybrid approach to implement template security has been proposed by the author A. Ghany et al. [13]. They have combined the feature transformation and biometric cryptosystem to increase the security of system. Principal Curves Approach algorithm is used to extract the features and random projection is performed as a transformation to project the original template. Also, by making use of k-means, Class Distribution Preserving transforms has been used to enhance the cancelable template into a binary template.
2. The author Nandakumar et al. [7] have proposed a method that uses passwords to harden the fingerprint fuzzy vault. In this scheme author have proposed various limitations of using only fuzzy vault. The vault becomes vulnerable if same biometric is used in another vault. This type of attack is known as correlation attack or cross-matching attack. Due to increase in number of chaff points FAR also increases. The proposed scheme enhances user privacy and prevents from cross-matching. Biometric template is transformed by using some random transformation and this transformed template is secured using fuzzy vault.
3. Few vulnerabilities in the work done by K. Nandakumar in his scheme [7] have been analyzed by the authors S. Hong et al. [14] and they have proposed a more

Table 1 Comparison of various fuzzy vault schemes

Name of author/year	Modalities used	Hybrid approach	Database used	Methodology used	Properties
Clancy et al./ 2003 [5]	Fingerprint	No	N.A.	Biometric template is directly secured using fuzzy vault.	Correlation attacks may be possible (only fuzzy vault used)
Uludag et al./ 2005 [6]	Fingerprint	No	IBM-GTDB	Biometric template is directly secured using fuzzy vault.	Correlation attacks may be possible (only fuzzy vault used)
Nandakumar et al./2007 [7]	Fingerprint	Yes	FVC2002–DB2 & MSU-DBI	Invertible transformation is applied on biometric template with key and transformed template is secured using fuzzy vault.	1. Prevention from correlation attacks (use of hybrid approach) 2. Revocability is possible
A.Ghany et al./2012 [13]	Fingerprint	Yes	N.A.	Biometric templates are secured by using transformation and bio- cryptosystem approaches together.	1. Prevention from correlation attacks (use of hybrid approach) 2. Revocability is possible
Hong et al./ 2008 [14]	Fingerprint	Yes	N.A.	Biometric template is transformed using a one way function which is then secured using fuzzy vault.	1. Prevention from correlation attacks (use of hybrid approach) 2. Revocability is possible
K. Nandakumar et al./2008 [10]	Fingerprint & Iris	No	MSU-DBI & CASIA	A key is used to transformed iris template. Combination of this key with fingerprint feature points is secured using fuzzy vault.	Prevention from correlation attacks (use of multimodal approach)
V.S. Meenakshi et al./2010 [11]	Iris & Retina	Yes	DRIVE & CUHK	Features are extracted from iris and retina. Invertible transformation is applied on these features and transformed features are secured using fuzzy vault.	1. Prevention from correlation attacks (use of multimodal and hybrid approach) 2. Revocability is possible
Li Yuan et al./2015 [12]	Face & Ear	No	FERET & USTB	Non-invertible transformations on ear and face to convert ordered set to unordered set. This set is secured using fuzzy vault.	Prevention from correlation attacks (use of multimodal approach)

secure scheme which can prevent security attacks on fuzzy vault. They have used one-way hash function, i.e., non-invertible function to transform the template and transform template is secured using fuzzy vault (Table 1).

5 Conclusion

In this paper we have reviewed the techniques for enhancing the security of fuzzy vault. Though fuzzy vault itself provides a very efficient way to secure biometric templates, but is still vulnerable to correlation attacks. Therefore by creating a hybrid system for biometric security we can increase the resistance of a fuzzy vault against such attacks.

References

1. Jain, A., Ross, A., Pankanti, S.: Biometric: a tool for information security. IEEE Transactions on Information Forensics and Security. 1, 125–143 (2006).
2. Jain, A., Nandakumar, K., Nagar, A.: Biometric Template Security. EURASIP Journal on Advances in Signal Processing. 2008, 1, 579416 (2008).
3. Rathgeb, C., Uhl, A.: A survey on biometric cryptosystems and cancelable biometrics. EURASIP J Inform Secur. 2011, 3 (2011).
4. Juels, A., Sudan, M.: A Fuzzy Vault Scheme. IEEE International Symposium on Information Theory. p. 408. IEEE, Lausanne, Switzerland (2002).
5. Clancy, T., Kiyavash, N., Lin, D.: Secure Smartcard-Based Fingerprint Authentication. ACM SIGMM Multim., Biom. Met. & App. pp. 45–52. ACM SIGMM (2003).
6. Uludag, U., Pankanti, S., Jain, A.: Fuzzy vault for fingerprints. International conference on Audio video based person authentication. pp. 310–319 (2005).
7. Nandakumar, K., Nagar, A., Jain, A.: Hardening Fingerprint Fuzzy Vault Using Password. ICB. pp. 927–937. LNCS 4642 (2007).
8. Ross, A., Jain, A., Nandakumar, K.: Handbook of Multibiometrics. Springer Science + Business Media, LLC, Boston, MA (2006).
9. Ross, A.: Introduction to Multibiometrics. 15th European Conference (EUSIPCO). Poznan, Poland (2007).
10. Nandakumar, K., Jain, A.: Multibiometric template security using fuzzy vault. IEEE Int. Conf. Biometrics: Theory, Applications and Systems. pp. 1–6. IEEE, Arlington, VA (2008).
11. Meenakshi, V., Padmavathi, D.: Security analysis of password hardened multimodal biometric fuzzy vault with combined feature points extracted from fingerprint, iris and retina for high security applications. Procedia Computer Science. 2, 195–206 (2010).
12. YUAN, L., LI, W.: Multimodal Template Protection Based on Data Transformation and Fuzzy Vault. Journal of Computational Information Systems. 11, 3999–4008 (2015).
13. Ghany, K., Hefny, H., Hassanien, A., Ghali, N.: A hybrid approach for biometric template security. IEEE International Conference on Advances in Social Networks Analysis and Mining. pp. 941–942. IEEE, Istanbul (2012).
14. Hong, S., Jeon, W., Kim, S., Won, D., Park, C.: The vulnerabilities analysis of fuzzy vault using password. IEEE, Second International Conference on Future Generation Communication and Networking. pp. 76–83. IEEE, Hainan Island (2008).

An Efficient Vector Quantization Based Watermarking Method for Image Integrity Authentication

Archana Tiwari and Manisha Sharma

Abstract This paper presents a two-stage watermarking technique for image authentication adapting advantages of vector quantization (VQ). In the present algorithm robust watermark and semifragile watermark are embedded independently in VQ compressed image in two successive stages. Robust watermark and VQ enhances the security of the system by providing double protection to designed system. A quantitative threshold approach using pixel surrounding error pixel is suggested for identification of attacks as acceptable or malicious. Experimental results demonstrate the capabilities of the method in classifying attacks and correctly locating tamper location. It is possible to detect and determine tamper with very high sensitivity. Present scheme outperforms previous algorithms and can distinguish malicious tampering from acceptable changes, and tampered regions are localized accurately.

Keywords Image authentication • Vector quantization • Watermarking • Semi-fragile watermarking • Attack classification • Tamper detection • Tamper localization

1 Introduction

Digital images play a significant role in almost all practical applications like military, medical, and broadcasting images. It is very easy to modify or manipulate digital images with advanced image editing software. Therefore, image authentication has therefore become a significant research topic. As per the recent study, digital watermarking [1, 2, 4, 6, 14, 17] is considered as the most suitable technique

A. Tiwari (✉)
Chhatrapati Shivaji Institute of Technology, Durg, India
e-mail: archanatiwari@csitdurg.in

M. Sharma
Bhilai Institute of Technology, Durg, India
e-mail: manishasharma1@rediffmail.com

P.K. Sa et al. (eds.), *Progress in Intelligent Computing Techniques: Theory, Practice, and Applications*, Advances in Intelligent Systems and Computing 719, DOI 10.1007/978-981-10-3376-6_24

for image authentication. Semi-fragile watermark is robust to changes which preserve contents of the image, while fragile to content altering modifications such as addition or deletion of an object, so it is suitable to for practical applications [5]. Vector quantization (VQ) [2, 3, 8], is broadly accepted in image compression applications [7, 9, 10, 13, 19] due to its high compression ratio and is very simple to decode [15]. VQ is an effective method in digital watermarking too. Various VQ techniques are recommended by researchers in past a decade for image watermarking [12, 14, 16, 18]; with the objective of image authentication. However, most of them are having poor visual quality issues and are not able to present quantitative methods for attacks identification. The concept of image watermarking using vector quantization technique was pioneered by [10]. Later modification of the paper was suggested in [11]. In [16] an image authentication algorithm fragile watermarking technique was proposed in the year 2016 [16], this method cannot tolerate any content preserving attacks due to fragile nature of the algorithm.

In proposed paper a new vector quantization-based watermarking method is suggested. The watermark is embedded in VQ compressed image; in two successive stages. In the first phase robust watermark is embedded for enhancing the security of image. In the second step, the semi-fragile watermark is embedded for image authentication. The proposed scheme further suggests a quantitative threshold-based approach for classifying attacks and localizing tampered area. The significant contributions of the present paper are:

- A VQ-based image authentication algorithm is designed where the watermark is embedded in VQ compressed image in two successive stages independently.
- The robust watermark is used to enhance the security of the scheme.
- Random keys are used in embedding robust and semi-fragile watermark, respectively; random nature of key improves the safety of designed system.
- The threshold-based approach is suggested to classify attacks as the acceptable or malicious.
- Tamper detection and localization is possible for a single pixel change. Thus, the sensitivity of the system is very high.

The present paper is organized as follows: Sect. 2 describes proposed image authentication algorithm, Sect. 3 gives Experimental results and performance analysis of image authentication algorithm, and finally conclusion of the paper is addressed.

2 Proposed Image Authentication Algorithm

2.1 Vector Quantization

The VQ method [8] is commonly known as the method for image compression. VQ process is performed in two stages, i.e., encoding and decoding. At VQ encoder section, the input image is partitioned into blocks, the index of code vector closest to block is assigned to all blocks. These indices are later used to reconstruct image at VQ decoder section. The proposed system consists of five constituents: Watermark embedding, watermark extraction, image authentication, and attacks classification.

2.2 Steps for Embedding Watermark

In proposed method watermarks are embedded in VQ compressed image in two successive stages. In first phase robust watermark is embedded to ensure the security of algorithm and in the second phase, the semi-fragile watermark is used for authentication purpose. Detail process is as follows:

Step 1—Partitioning of training image: The original image $S_i, (i = 1, 2, 3, \ldots, 512)$; is partitioned into 16,384 non-overlapping blocks, of size 4 × 4. Input training vectors for first stage is obtained from original image $S = \{x_i \in R_d | i = 1, 2, \ldots, 512\}$. Robust watermark bit sequence W_R is permutated using passkey1, where $R = 1, 2, 3, \ldots, 128$.
Step 2—Codebook1 generation: To generate codebook1, $C1 = \{c_j \in R_d | j = 1, 2, \ldots, 256\}$ following process is followed. Initially distortion (D) is set as $D_0 = 0$ and iterations (k) is set as $k = 0$. Input n vectors are classified into K clusters according to $x_i \in S_q$ if $\|x_i - c_q\|p \leq \|x_i - c_j\|p$ for $j \neq$ q. Then cluster centers, is updated as c_j, where c_j, j = 1, 2, 3, …, 256 and defined as $C_j = 1/|S_j| \sum x_i \in S_j x_i$, where S is original image and X is partitioned image. Set $k \leftarrow k + 1$ and then distortion is computed, the distortion $Dk = \sum_{j=1}^{k=j} \sum_{x_i S_j} \|x_i - c_j\|p$. If $(Dk - 1 - Dk)/Dk > Q$ (a small number), then the process is repeated. Finally codebook C is obtained as,$C = \{c_j \in R_d | j = 1, 2, 3, \ldots, 256\}$.
Step 3—VQ encoding initial stage: In VQ encoder, each vector of input training vector searches for its best code vector in the codebook for each input image block the nearest code vector is computed. The number of squared Euclidean distances is equal to 'i', i.e., size of the codebook. Then, the closest codeword is calculated by finding the minimum squared Euclidean distance from itself to the original image block [5]. The binary index of the selected codeword is sent to the decoder. The decoder has the same codebook and can get back the codeword from given the binary index. The VQ compressed image, i.e., $Z(m, n)$ is reconstructed using these, indices i, i = 1, 2, 3, …, 256.
Step 4—First stage watermark embedding: For watermark embedding in the first phase, the threshold is set which is equal to half codebook size. Polarity is computed using variance; it is set to one if it is greater than threshold else it is reset to zero. Embedded robust watermark is generated using $W_1 = W_R \oplus polarity$, where W_R is the robust watermark. First stage watermarking does not affect VQ compressed image; this phase provides security to designed algorithm.
Step 5—Codebook1 generation second stage: The difference between first stage input and output is calculated which is quantization error (Q) of first stage

$Q = S(m, n) - Z(m, n)$. For the second phase, codebook C_2 is generated using the method described in step 2; here partitioned error image (Q) is used in place of S_i, i.e., original image for codebook generation. After codebook generation, each vector of training vector finds its best code vector in codebook C_2 and indices $i, j = 1, 2, 3, \ldots, 256$, is assigned to $Q(m, n)$.

Step 6—Second stage watermark embedding: This error image $Q(m, n)$, is partitioned into 16,384 non-overlapping blocks, where Q_i, (i = 1, 2, 3, …, 512). Individual blocks of size 4 × 4 are used to embed semi fragile watermark, W_s for $s = 1, 2, 3, \ldots, 128$. A single bit is inserted in each index in the second stage using passkey2, which is used to assign embedding positions for adding a single bit of semi-fragile watermark in each index. Now modified indices are used to rebuild semi fragile watermarked image, i.e., $Q^{'}(m, n)$. Here passkey2 is randomly generated, so each, time algorithm is executed; key positions are different, this improves the security of algorithm.

Step 7—Watermarked image construction: Finally, watermarked image is obtained by combining first-stage and second-stage outputs. Thus watermarked image is, $W(m, n) = Z(m, n) + Q^{'}(m, n)$, where $m = n = 512$.

2.3 Steps for Watermark Extraction

Once test image, i.e., watermarked image $W^{'}(m, n)$ is received; the watermark is extracted from it using inverse process. The novelty of scheme is the watermark and is embedded independently in both stages, and random keys are used for watermark embedding, so unique keys are used for each training image.

Step 1—Partitioning watermarked image: In receiver side reverse process of watermark insertion procedure is carried out, the received image $W^{'}(m, n)$ is further partitioned into non-overlapping blocks of size 4 × 4. Simple VQ encoder is used along with codebook1 to find encoded indices; further, these indices are used to reconstruct the first stage o/p.

Step 2—Robust watermark extraction: In the first phase of robust watermark extraction, codebook1 and VQ decoder is used to extract encoded indices of watermarked image. Encoded indices are obtained by searching nearest neighbourhood indices. Polarity is computed from indices of stage 1, using standard deviation. Then XOR operation is performed between polarity and passkey1 to obtain permuted robust watermark, the final robust watermark is obtained by the reverse process.

Step 3—Semifragile watermark extraction: In the second stage, watermarked image is segmented using the first phase o/p. Passkey2 is used to find the watermarking position in the segmented image. These watermarking bits are grouped to form semi fragile watermark W_s.

2.4 Steps for Image Authentication and Attack Classification

After successful extraction of the watermark in two stages, each stage's o/p are compared with embedded watermark to find similarity. The received image is further tested for the possible attack using threshold-based approach.

Step 1—Imperceptibility and embedding capacity: The imperceptibility is measured from PSNR (peak to signal ratio), the higher value of PSNR shows better imperceptibility.

$$\mathrm{PSNR} = 10\log_{10} 10\left(\frac{255\times 255}{\mathrm{MSE}}\right)\mathrm{dB} \tag{1}$$

$$\mathrm{MSE} = \frac{1}{I\times J}\sum_{m=0}^{m=I-1}\sum_{n=0}^{n=J-1}\left(S(m,n)-W(m,n)\right)^2, \tag{2}$$

where $I = J = 512$ and $m = n = 512$. Mean square error (MSE) and PSNR is computed between the original image and watermarked image using Eqs. 1 and 2. The maximum embedding capacity (MS) of the watermarking method [19] is calculated using Eq. 3.

$$\mathrm{MS} = \frac{M\times N}{a\times a} \tag{3}$$

where $M \times N$ is the size of cover image and $a \times a$ is the size of watermark image size.

Step 2—Similarity check between embedded and extracted watermark: Similarity between inserted and extracted watermark is calculated using normalized hamming similarity, i.e., NHS [16]. NHS is defined as

$$\mathrm{NHS} = 1 - \frac{\mathrm{HD}(X,W)}{128\times 128} \tag{4}$$

where HD(.) denotes the Hamming distance between two binary images, i.e., the number of different bits of the two binary images [16]. A closer value of NHS to one means no distortion.

Step 3—Error pixel identification: If $\mathrm{NHS} \le 0.099$, then to find error pixels in received image $W'(m,n)$, the test image $(W'(m,n))$ is *XORed* with received image $W(m,n)$. $R(m,n) = W(m,n)\,XORW'(m,n)$, if any pixel of $R(m,n)$ is 1 it is considered as error pixel. Otherwise, if the pixel value is zero, it is not an error pixel.

Step 4—Calculation of percentage of error pixel: After identifying error pixels in difference image $R(m,n)$. Each error pixel is checked by using 3 × 3 pixel neighbourhood approach. A pixel is considered as malicious pixel if more than four pixels of its surrounding eight neighbourhood pixels are having value one otherwise it is not a malicious pixel.

Let $N1 =$ No of pixels having value one in $R(m, n)$,
$N =$ Total no of pixels in $R(m, n)$
$N2 =$ No of pixels having more four error pixel in its eight surrounding neighbourhood.
$T =$ Total no of error pixels in $R(m, n)$, then

$$T1 = (N1/N) \tag{5}$$

$$T2 = (N2/T) \tag{6}$$

Step 5—Thresholding for attack classification: Thus, present scheme uses threshold-based check to classify attacks. Algorithmic structure of proposed system is

```
Compute N1 from image $R(m,n)$.
If T1=0, then image is not attacked. The image is considered
as authentic.
      End if
      T1≠0, then compute the value of NHS.
If NHS> 0.7, then extracted image is recognizable.  Further,
check for authenticity of the image.
If  T2<T,  the  image  is  incidentally  attacked,  it  is
acceptable manipulation. Thus, the image is authentic.
      Else
If  NHS<  0.7,  T1>0  and  T2>T,  then  image  is  maliciously
attacked, it is not authentic.
      Then  tamper  is  detected  and  localized  using
      image $R(m,n)$.
```

Step 6—Tamper detection and localization: If $T2$ is more than threshold and NHS are less than 0.7, then the image has maliciously tampered.

3 Experimental Result and Performance Analysis

The performance of the proposed algorithm is assessed using MATLAB 10 software; the experiment is conducted on 150, 8-bit images of size 512 × 512. These images are further divided into 16,384 blocks of size 4 × 4. Binary watermark images of size 128 × 128 are embedded in both stages. The codebook is obtained using standard LBG algorithm as mentioned in step 2 of watermark embedding subsection.

3.1 Codebook Selection and Embedding Capacity

Table 1 shows simulation results for different combinations of the codebook here values in rows shows codebook1 × codebook2. The table shows analysis of watermarked image quality for various codebook sizes. It can be inferred that better watermarked image quality is obtained when both codebooks are of size 256 × 256. Therefore, proposed work uses codebooks of size 256 × 256.

In proposed technique, robust watermark and the semifragile watermark can be embedded following steps 3–7 of watermark embedding subsection. Embedding capacity of the proposed scheme is calculated using Eq. 3.

$MS = (512 \times 512)/(128 \times 128) = 16$, i.e., 1 bit of watermark can be embedded in a block of size 4 × 4.

3.2 Robustness to Common Image Processing Manipulations

Image processing attacks such as blurring, low-pass filtering using Gaussian filter, salt and pepper noise, rotation at different angles and JPEG compression attacks are performed on 120 watermarked images. Tables from 2, 3 to 4 demonstrate the result of only 4 test images. The nomenclature used in tables are, T_w shows error pixels obtained from extracted watermark image, T is percentage no of error pixels in difference image $R(m,n)$, $T2$ is the percentage of malicious pixels in difference image $R(m,n)$. It is observed that extracted image is recognizable when

Table 1 PSNR values of watermarked image from different-sized codebooks

Codebook image	256 × 256 PSNR	256 × 512 PSNR	128 × 512 PSNR	16 × 256 PSNR
Peppers	40.1861	38.1564	39.8423	40.4231
F-16	41.7308	37.5834	37.8911	38.1272
Lena	41.7667	39.9176	37.5134	39.8133
Cameraman	41.5551	40.8254	40.7912	40.0845
Baboon	40.8900	37.6192	38.9943	40.3731

Table 2 Simulation results of watermark ed test images for Blur attack

Image	T_w	$\%T$	$\%\ T2$	PSNR	NHS (semi)	Parameter	Classification
Pepper	305	52.4	52.3	38.49	0.8112	Radius = 1	Incidental
Baboon	476	79.9	50.06	35.56	0.7067	Radius = 1.1	Incidental
Lena	326	87.7	55.12	38.86	0.8082	Radius = 1.1	Incidental
F-16	234	56.3	41.24	37.28	0.857	Radius = 1	Incidental

NHS > 0.07 and threshold value for the image is set as 57% based on experimental results.

Figure 1 gives the visual presentation of robustness characteristics of the scheme for different images. From experimental results, it can be inferred that present algorithm can clearly distinguish between malicious and incidental attacks. The algorithm outperforms most of the recent algorithms too [2, 5, 13, 16, 19] concerning image quality, attack classification and robustness for content preserving manipulations. The extraction of the watermark is done independently in two stages using different codebooks, i.e., codebook1 for the first phase and codebook2 for the second stage; this resulted in improved quality of extracted watermark.

Table 3 Simulation results of watermarked test images for Gaussian filter (3 × 3)

Image	T_w	$\%T$	$\% T2$	PSNR	NHS (semi)	Parameter	Classification
Pepper	303	71.67	54.1	44.31	0.8163	Sigma = 0.8	Incidental
Baboon	477	80.13	45.81	36.49	0.9136	Sigma = 0.6	Incidental
Lena	815	75.48	51.8	40.46	0.9503	Sigma = 0.6	Incidental
F-16	749	67.06	37.37	37.68	0.9543	Sigma = 0.6	Incidental

Table 4 Simulation results of watermarked test images for salt and pepper noise

Image	T_w	$\%T$	$\% T2$	PSNR	NHS (semi)	Noise density	Classification
Pepper	4986	6.97	0	41.92	0.7136	0.07	Incidental
Baboon	4712	6.94	0	42.52	0.7075	0.07	Incidental
Lena	4924	6.04	0	48.76	0.7679	0.06	Incidental
F-16	4881	6.97	0.005	40.18	0.7021	0.07	Incidental

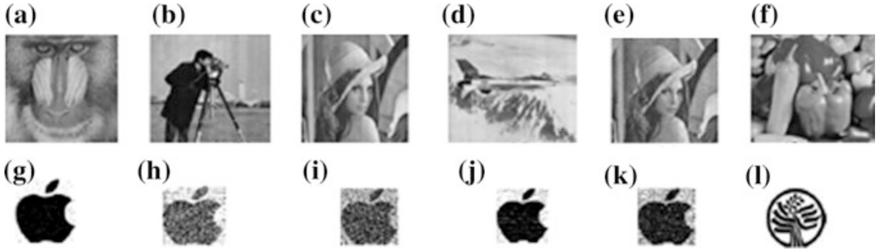


Fig. 1 Illustration of robustness nature of present scheme figure shows, **a** Gaussian filtered (3 × 3 sigma = 0.6), **b** Blurred image radius = 2. **c** Rotation at 0.2 clockwise. **d** jpeg compression at $Q = 70$. **e** Salt and pepper at 0.07. **f** Watermarked image under no attack; figures **g–k** show extracted semi-fragile watermark of respective attacked images above them, and figure **l** shows extracted robust watermark for attacked images

3.3 Fragileness Characteristics for Malicious Attacks

Different malicious attacks are performed on 120 watermarked images to show the effectiveness of scheme in detecting and localizing malicious attacks. Three kinds of malicious attacks are applied on watermarked image, like cutting/substituting another object, cropping a part of the image and inserting a text in watermarked image. As shown in Fig. 2 (Table 5).

$$\text{Tampering ratio} = (\text{window} * \text{window}) / \text{total no of pixel in input image}$$

It can be inferred from above table that proposed algorithm gives better results in comparison to other algorithms both in PSNR and similarity factor.

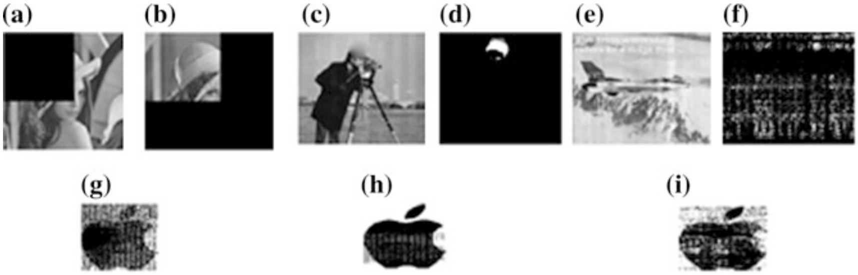


Fig. 2 Illustration of malicious attacks: **a** crop attack and **b** difference image and figure **g** shows extracted semi-fragile watermark, **c** shows cut attack, **d** shows difference image, **h** shows extracted semi-fragile watermark, **e** text attack, **f** shows corresponding difference image, **i** shows extracted semi-fragile image

Table 5 Comparative performance of different algorithm for detection capacity

Reference no.	Tampering ratio in %	PSNR	Similarity	Tampering ratio in %	PSNR	Similarity
Proposed	10	38.96	0.981	25	34.67	0.8853
[19]	50	…	0.497	25	…	0.749
[11]	–	–	–	25	29	0.948
[20]	9.7	35.23	- - -	25	34.43	- - -
[16]	10	36.21	0.976	20	33.29	0.954

4 Conclusions

A novel vector quantization-based semi-fragile watermarking scheme for image authentication application is presented using two stages of watermark embedding technique; in the first phase robust watermark is embedded while in the second stage semi-fragile watermark is embedded using modified index method. The novelty of proposed work is in randomness in embedding key generation, independent watermark embedding in two stages; and threshold-based quantitative analysis is suggested for attacks classifications. Simulation results show that the vector quantization based watermarking method outperforms most of the previous algorithm regarding watermarked image visual quality, robustness nature, tamper detection, and localization. The algorithm is tested for different acceptable and malicious attack. Certain standard image processing manipulations, such as noise addition, rotation at different angles, JPEG compression and filtering are analysed. The proposed algorithm can identify cut, crop, substitute and text addition attacks as malicious attacks. Moreover proposed work shows high sensitivity, it can detect a single pixel change in received image and can classify tampered area as malicious for window size is 8 × 8. Thus experimental result shows the effectiveness of vector quantization based watermarking model; which provides insights for authentication of the image and allows better control of robustness and fragility to content altering attacks and excellent tamper localization capacity. In future proposed work will focus on content recovery from maliciously tampered image.

References

1. Barreto, P. S., Kim, H. Y., & Rijmen: Toward secure public-key block-wise fragile authentication watermarking, in IEE Proc. Vision, Image and Signal Processing, 149(2): 57–62, (2002).
2. Chuang, Jun-Chou, and Yu-Chen Hu: An adaptive image authentication scheme for vector quantization compressed image, Journal of Visual Communication and Image Representation 22(5):440–449, (2011).
3. Cox, I. J., & Miller, M. L.: Review of Watermarking and The Importance of Perceptual Modelling, Electronic Imaging, 92–99, (1997).
4. Khalil MS, Kurniawan F, Khan MK, Alginahi YM.: Two-layer fragile watermarking method secured with chaotic map for authentication of digital holy Quran. The Scientific World Journal, (2014).
5. Kuarto Maeno: New semifragile image authentication techniques using random bias and nonuniform quantization, IEEE Transactions on Multimedia, 8(1): 32–45, (2006).
6. Li, M., Xiao, D., & Zhang, Y: Attack and improvement of the fidelity preserved fragile watermarking of digital images, Arabian Journal for Science and Engineering, pp. 1–10, (2015).
7. Lin, Chia-Chen, Yuehong Huang, and Wei-Liang Tai.: A novel hybrid image authentication scheme based on absolute moment block truncation coding, Multimedia Tools and Applications, pp. 21–26, (2015).
8. Linde, Y., Buzo, A., & Gray, R. M: An algorithm for vector quantizers design, IEEE Transactions on Communications, 28(1), 84–95, (1980).

9. Iliyasu, A.M., Le, P.Q., Dong, F. and Hirota, K: Watermarking and authentication of quantum images based on restricted geometric transformations. *Information Sciences*, *186*(1), 126–149, (2012).
10. Lu, Z. M. & Sun S. H.: Digital image watermarking technique based on vector quantization, in Electronics Letters, 36(4):303–305, 2000.
11. Lu, Z. M., Liu, C. H., Xu, D. G., & Sun, S. H: Multipurpose image watermarking algorithm based on multistage vector quantization, IEEE Trans. on Image Processing, 14 (6): 822–831 (2005).
12. Lu, Z. M., Liu, C. H., Xu, D. G., & Sun, S. H: Semi-fragile image watermarking method based on index constrained vector quantisation, in Electronics Letters, 39(1):35–36, (2003).
13. Makur A. & Selvi S. S: Variable dimension vector quantization based image watermarking, Signal Processing, 81(4): 889–893, (2001).
14. Miller, M. L., Cox, I. J., Linnartz, J. P. M., & Kalker, T: A review of watermarking principles and practices, Digital Signal Processing for Multimedia Systems, pp. 461–485, (1999).
15. Ning C. H. E. N. & Jie, Z. H. U., Multipurpose speech watermarking based on multistage vector quantization of linear prediction coefficients, China Universities of Posts and Telecommunications, 14(4):64–69, (2007).
16. Qin, C., Ji, P., Wang, J., & Chang, C. C.: Fragile image watermarking scheme based on VQ index sharing and self-embedding, Multimedia Tools and Applications, pp. 1–21, (2016).
17. Rosales-Roldan, L., Cedillo-Hernandez, M., Nakano-Miyatake, M., Perez-Meana, H., & Kurkoski, B: Watermarking based image authentication with recovery capability using halftoning technique. Signal Processing: Image Communication, 28(1), 69–83, (2013).
18. Shen, Jau-Ji, & Jia-Min Ren: A robust associative watermarking technique based on vector quantization, Digital Signal Processing 20(5): 1408–1423, (2010).
19. Wu, H. C. & Chang C. C.: A novel digital image watermarking scheme based on the vector quantization technique, Computers & Security, 24(6):460–471, (2005).
20. Yang, Chun-Wei, and Jau-Ji Shen: Recover the tampered based on VQ indexing, Signal Processing 90(1), pp. 331–343, (2010).

XSS Attack Prevention Using DOM-Based Filter

Asish Kumar Dalai, Shende Dinesh Ankush and Sanjay Kumar Jena

Abstract Cross-site scripting (XSS) is one of the most critical vulnerabilities found in web applications. XSS vulnerability present in web application that takes untrusted data and sends it to a web browser without proper input validation. XSS attack allows the adversary to execute scripts in the victim browser which can deface web sites, hijack user sessions, or redirect the user to malicious contents. Some of the proposed methods to XSS attack include the use of regular expressions to identify the presence of malicious content. However, this can be bypassed using parsing quirks and client-side filtering mechanisms such as Noscript and Noxes tool. The existing solutions are comparatively slow and cannot withstand against all attack vectors. Some of the existing approaches are too restrictive resulting in loss of functionality. In this paper, an API for server-side response filtering has been developed. The proposed method allows the HTML to pass through but blocks the harmful scripts. Unlike other approaches it requires a minor modification in existing web application. The performance evaluation shows that the proposed technique is having high fidelity and comparatively less response time.

Keywords XSS attack · Input validation · DOM · Web application security

A.K. Dalai (✉) · S.K. Jena
National Institute of Technology Rourkela, Rourkela, India
e-mail: dalai.asish@gmail.com

S.K. Jena
e-mail: skjena@nitrkl.ac.in

S.D. Ankush
Maverick Labs Pvt. Ltd., Pune, India
e-mail: dineshshe@gmail.com

P.K. Sa et al. (eds.), *Progress in Intelligent Computing Techniques: Theory, Practice, and Applications*, Advances in Intelligent Systems and Computing 719, DOI 10.1007/978-981-10-3376-6_25

1 Introduction

Cross-site scripting (XSS) is a class of attack that targets the application layer of the TCP/IP network stack. XSS usually targets scripts embedded in a web page that is executed on the client side (in the client's web browser/application) rather than on the server side. XSS occurs due to the security flaws of several client-side technologies such as VBScript, JavaScript, HTML, Flash, and ActiveX, etc. The existence of weakness in these technologies is the prime cause of the exploit. XSS is initiated by modifying the client-side scripts that present in the web application in a way as desired by the adversary.

1.1 Types of XSS Attacks

Based on the method followed for the accomplishment of XSS attack, they can be broadly categorized into the following classes.

- Non-persistent/Reflected XSS Attack: In non-persistent attacks the user-supplied data is directly reflected back to the user in the form of error message, search result, welcome message, or any other type of server response. Such kind of vulnerability can be exploited using social engineering attack, where the attackers convince the victim to click a malicious link that is vulnerable to XSS.
- Persistent/Stored XSS attack: Persistent or stored XSS is a kind of attack where malicious scripts injected by the attacker are stored permanently in the server either in the web pages or database. Clients who access that vulnerable web page became a victim of XSS attack. The clients treat as a normal page from trusted web application and interpret that without any proper HTML validation and eventually the malicious code incorporated in that vulnerable page gets executed at the client side.
- DOM-based XSS attack: In the case of DOM-based XSS attack, the payload is executed as a result of alteration of the DOM environment in the user web browser. In such attack, it modifies the original client-side script so that the client-side code runs in an undesired manner. In DOM-based XSS, the malformed script is not sent to the web server. DOM-based XSS attack takes place when user-supplied data is interpreted as JavaScript using methods such as *innerHTML*, *document.write*, or *eval()*.

1.2 Impact of XSS Attack

The severity of XSS attack may cause session hijacking by cookie stealing to get session ID that can be used impersonate as authorized user. The attack can also allow to access the browsing history of the victim which compromises the privacy of the victim.

2 Literature Survey

Researchers have proposed several client-side and server-side XSS prevention techniques in the past. These techniques involve modification in a web application or alteration at the browser of the client.

2.1 Server-Side Mitigation

There exist several XSS attack mitigation techniques that can be implemented at client side or server side. HTML Purifier [1], AntiSamy [2] libraries are some of the existing solution that uses input validation and blacklisting to prevent XSS. Some of existing solutions [3, 4] are based on server-side modification. But these modifications are complex enough and generates a lot of burden for the developer to deploy those changes to the code and also results in increased response time. The statistical analysis method used by deDacota [5] technique prevents XSS attack by rewriting the web application completely to remove the inline JavaScript.

2.2 Client-Side Mitigation

For mitigation of XSS at the client side, Noxes [6] is an application-level firewall that carries all Web traffic using client-side web proxy. Noxes [6] requires user-specific settings unlike SWAP [7] and it also if there occurs any new event that is not present in the firewall rule it needs the user interaction. Another tool named XSS Auditor [8] creates the interface between the JavaScript engine and HTML parser of the browser and it has high performance and high reliability. Google Chrome has by default enabled its implementation in the browser.

3 Threat Model

We need to consider the types of attacks that need to be handled before designing any filter. In this section, we have discussed the issues that can lead to the XSS vulnerability.

3.1 HTML5 Vulnerable Features

XSS vulnerability may be caused by some of the newly added features such as 'autofocus,' 'formaction,' 'attribute,' and 'srcdoc' attribute of <iframe> tag.

3.2 Parsing Quirks

A different browser parses the comments in a different way. If the user-supplied data contains comments, it could create a problem.

3.3 Event Attributes

On the occurrence of a particular event, the value associated with the event attributes comes to action. This can be used to access DOM properties or execute the malicious script.

3.4 Vulnerable DOM Properties

To access numerous entities of the document, the browser uses the DOM properties. As it can access any tag and their attribute from the DOM object it can obtain the history of the document and also the user cookie, which reveals the browsing behavior and the sensitive information about the user.

3.5 Data URIs

For saving the fetch time, data URIs are used where data pointed by the external source can be combined within it as self-content entities. Therefore, data URIs may be used as a medium to exploit the XSS vulnerability by including JavaScript in them.

3.6 Encoded Attribute Values

Several encoding methods such as decimal/hexadecimal HTML character encoding, URL encoding, Hex encoding, base 64 encoding, HTML entity encoding, etc. are being used by the attacker to evade the filters or input validation mechanism based on regular expressions.

3.7 Cascaded Style Sheet Vectors

Attackers can use style sheet properties like "-o-link and -o-link" source which allows scripts as its values. For supplying CSS information and to display images browsers like Internet Explorer support these features.

3.8 History Tampering

The APIs like *history.pushState()* and *history.replaceState()* allow to change and create the users history. This can be used for phishing attacks or obfuscate bad intentions by an attacker. The attacker can change the address bar information and the location DOM object.

3.9 HTML Tag and Attributes

Some HTML tag and their attributes can be used to redirect the user to an external website. Therefore, an attacker may use these tags to point it to malicious contents.

4 Problem Statement

Unlike existing approaches, the proposed solution should not burden the client. The XSS attack prevention technique should not require large changes in the existing applications. The proposed solution should not delay the response. So, there exists a trade-off between security achieved and the response time.

Blocking the malicious script injection and identifying novel attack vectors without hampering the functionality of the application is a challenging task.

5 Proposed Method

In the proposed solution, at server side, we filter the responses, which contain user-supplied data which reflects back to the client to be rendered in his browser. There will be the boundary tag, which will separate all the user-supplied data in the server response from the total content of the web application. Therefore, from the server response, the user-supplied data are checked to for malicious scripts and are filtered if found. After filtering the data is embedded in the user response again.

DOM-based filtering method has been used to detect and remove the malicious scripts. To parse the user-supplied data in the DOM tree, we have used the HTML parser. The Dom tree is then traversed as given in the following steps.

1. First of all, the entire content is decoded and converted into its original form.
2. The managed content is parsed using Jsoup HTML parser to document object model.
3. The DOM is then examined for the presence of several event attributes. If any of these attributes are reading any vulnerable DOM properties, then those are filtered out from content.
4. Then filtered DOM is checked for attributes such as ‘formaction’ and if found are filtered out.

5. Then DOM is searched for vulnerable attributes like 'background' and checked for their values against the whitelist. If attribute values are not present, then they are removed from the output DOM.
6. Then filtered DOM is inspected for embedded CSS in <style> tag and attributes. If the CSS content has malicious scripts, then these CSS contents are removed.
7. Then the DOM is looked for <svg> tag, and if it points to any dynamic HTML content, then the attribute is blocked.
8. The filtered DOM is examined for a vulnerable combination of the tag attribute pair and their values are checked with allowable extensions. If present in the allowed file list, then kept else filtered.
9. The DOM is searched for the presence of data URIs. The values associated with them are decoded using appropriate decoding mechanism and checked for malicious content if found then removed.
10. Then resulted DOM is scanned for special tags such as <base > and <script> and managed accordingly.

6 Implementation and Results

The most robust parser like Jsoup HTML parser [9] has been used in the proposed method. The comment handling procedure of Jsoup has been modified as per the requirement. And also we have modified Jsoup to support boundary tag introduced by our method.

To filter the response from the server we used Java filter interfaces. The proposed method has been tested on a vulnerable web application developed using JSP hosted in the local XAMPP server.

6.1 Fidelity Results

A total of 230 different attack vectors have been taken from different sources such as HTML5 security cheat sheet [10], XSS cheat sheet [11] has been taken to validate the proposed model. Some of the attack vectors could not work due to the changes in the updated browsers. The result of the attack detection and filtering is shown in Table 1. It has been found that the proposed method can withstand all attack vectors tested in different browsers.

6.2 Response Time Analysis

To analyze the response time, we have used the Firefox web browser along with the XAMP server. The firebug add-on is used to calculate the response time. For each

Table 1 Attack detection and filtering statics of the proposed method

Browser	No. effective attack vectors	Attacks detected and filtered by proposed method	Undetected attack by proposed solution
Firefox 39	108	108	0
Chrome 44	106	106	0
IE 11	119	119	0
Opera 31	115	115	0
Safari 5.1.7	104	104	0

Table 2 Response time statics of the filter

Size KB (Approx.)	Response time w/o filter (in ms)	Response time with filter (in ms)	Difference in response time (in ms)	Response factor
2	8.6923	33.2307	24.5384	2.823
10	14.461538	55.08255	40.6210	2.8089
50	39.9580	151.7272	111.7692	2.7961
100	85.0019	306.6676	221.6657	2.6077
200	137.1838	501.1111	363.9273	2.6528

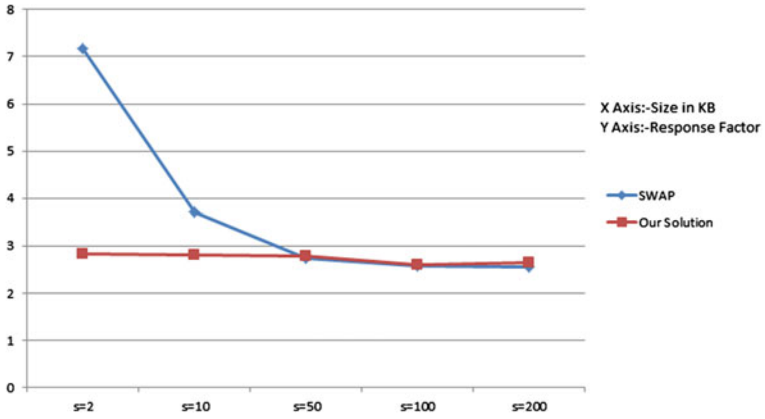


Fig. 1 Response time comparison

page of certain size 20 reloads have been done and the average time is calculated for analysis. Table 2 shows the detail statistics of response time.

The comparison results of the proposed method against SWAP [7] are shown in Fig. 1.

7 Conclusions and Future Work

Rather than using modified web browser as proposed by Wurzinger et al. the proposed technique uses an API to stop the malicious scripts from being executed. The results show that our method has less overhead than its counterparts. Unlike the method proposed by Chandra et al. our technique blocks the XSS attack for all popular web browsers instead of the one which is being used for malicious script detection.

As the proposed method is implemented on the server side it can only detect and block the server-side attacks but cannot detect the DOM-based XSS attacks. The proposed method uses whitelist-based approach, and all known XSS vulnerabilities present in JavaScript and HTML5 that are available to date are included in the list. A zero-day attack may bypass the proposed filtering mechanism, and there are further research scopes to address the zero-day attacks.

References

1. Html purifier. http://htmlpurifier.org/ (2016)
2. Antisamy. https://code.google.com/p/owaspantisamy (2016)
3. Kieyzun, A., Guo, P.J., Jayaraman, K., Ernst, M.D.: Automatic creation of sql injection and cross-site scripting attacks. In: Software Engineering, 2009. ICSE 2009. IEEE 31st International Conference on, IEEE (2009) 199–209
4. Bisht, P., Venkatakrishnan, V.: Xss-guard: precise dynamic prevention of cross-site scripting attacks. In: Detection of Intrusions and Malware, and Vulnerability Assessment, Springer (2008) 23–43
5. Doupé, A., Cui, W., Jakubowski, M.H., Peinado, M., Kruegel, C., Vigna, G.: dedacota: toward preventing server-side xss via automatic code and data separation. In: Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security, ACM (2013) 1205–1216
6. Kirda, E., Kruegel, C., Vigna, G., Jovanovic, N.: Noxes: a client-side solution for mitigating cross-site scripting attacks. In: Proceedings of the 2006 ACM symposium on Applied computing, ACM (2006) 330–337
7. Wurzinger, P., Platzer, C., Ludl, C., Kirda, E., Kruegel, C.: Swap: Mitigating xss attacks using a reverse proxy. In: Proceedings of the 2009 ICSE Workshop on Software Engineering for Secure Systems, IEEE Computer Society (2009) 33–39
8. Bates, D., Barth, A., Jackson, C.: Regular expressions considered harmful in client-side xss filters. In: Proceedings of the 19th international conference on World wide web, ACM (2010) 91–100
9. Jsoup html parser. http://jsoup.org/ (2016)
10. Html5 security cheatsheet. http://html5sec.org/ (2016)
11. Xss prevention cheatsheet. https://www.owasp.org/index.php/XSS_(Cross_Site_Scripting)_Prevention_Cheat_Sheet (2016)

Part IV
Big Data and Recommendation Systems

Friendship Recommendation System Using Topological Structure of Social Networks

Praveen Kumar and G. Ram Mohana Reddy

Abstract Popularity and importance of Recommendation System is being increased day by day in both commercial and research community. Social networks (SNs) like Facebook, Twitter, and LinkedIn draw more attention since without any previous knowledge a lot of connections have been established. The creation of relationship between users is the key feature of a social network. Therefore, it is important for researchers to look for a new way to provide recommendations with more relevance. This paper proposes two algorithms for recommending a new friend in online social networks. The first algorithm is based on the number of mutual friends and second is based on influence score. These recommendation algorithms use collaborative filtering and provide the idea of doing recommendations (e.g., Facebook recommend friends, Netflix suggest movies, Amazon recommend products, etc.). Obtained results and analysis indicate that influence-based recommendation system is more accurate as compared to mutual friend-based recommendation. These proposed recommendation algorithms can be used for the development of an effective social networking or e-commerce site and thereby providing a better experience to users.

Keywords Recommendation system · Social networks (SNs) · e-commerce · Facebook · Twitter

1 Introduction

Recommendation system has become popular in social network and e-commerce services [1] in order to reduce the search query generation and processing. It helps people make new friends in Social Networks or suggest related products to buy. As

P. Kumar (✉) · G.R. Mohana Reddy
National Institute of Technology Karnataka,
Surathkal, Mangalore 575025, Karnataka, India
e-mail: agrawalpraveen241@gmail.com

G.R. Mohana Reddy
e-mail: profgrmreddy@gmail.com

P.K. Sa et al. (eds.), *Progress in Intelligent Computing Techniques: Theory, Practice, and Applications*, Advances in Intelligent Systems and Computing 719, DOI 10.1007/978-981-10-3376-6_26

background, generally recommendation systems make a list of recommendation by using two approaches: through collaborative or content-based filtering.

Collaborative filtering methods predict what users will like based on their similarity with other users by collecting and analyzing a large amount of information on users preferences, activities, and behaviors. In simple words, it predicts the future behavior/preferences with similarity of some other users behavior/preferences [2] in the past. For e.g. suppose User1 likes A, B, and C and User2 likes A, B, C, and D then it stands for the reason that User1 will like D as well. This recommendation system does not consider any other information about the objects like their characteristics, color, size, etc. These all objects A, B, C, and D can be a person, product, place, or anything else.

Contents-based filtering methods consider the characteristics of objects we like and based on these characteristics it suggests similar sorts of objects. Keywords are used in content based filtering to describe the item and based on the type of item a user likes, a user profile is built [3]. This type of recommendation system recommends the best matching objects among various candidate objects which are similar to those that a user is liking at present or liked in the past.

For online social networks [4], recommendations can be done by utilizing two features: topological structure and non-topological structure (user profile) of a social network. Topology-based recommendation systems are mainly based on the connectivity of users, checks whether friend of friends of existing user can become new friend to that user. In this method, similarity between nodes are obtained by using intrinsic properties of network structure. Non-Topology-based recommendation system uses the other information (like interest, location, age, etc.) from user's profile.

A key issue with content-based filtering and recommendation system is that it is limited to recommend only similar type of objects which user has already used or using at present. This is having less significant when other objects from other environments can be recommended based on his friends' preference.

The core idea behind the proposed work is to use Collaborative filtering for friendship recommendation. First, we designed an algorithm considering friends-of-friends and mutual friends for recommendations. Second, we proposed an idea of influence in friendship recommendation by prioritizing the nodes or person having lesser friends and have greater influence on their friends for suggesting a new friend.

The Key Contributions of this paper are:

- Design and development of ***two*** Friendship Recommendation algorithms considering number of mutual friends and influence factor.
- A comparative study and analysis for accuracy of both algorithms on a large-scaled standard data set of Facebook and Twitter SNs.

The rest of this paper is organized as follows: Sect. 2 deals with the Related Work; Sect. 3 explains the Proposed Methodology; Sect. 4 deals with the Results and Analysis; finally concluding remarks and future work are included in Sect. 5.

2 Related Work

In this section, we review some of the earlier work including recommendation techniques and methods used in the area of recommendation system of online social networks. There are several existing works in this area, some of them incorporated the idea of user's reviews using collaborative filtering [5, 6], whereas others use social network structure for development of recommendation systems.

Garton et al. [7] described the various parameters (like content, location, strength, directions, etc.) that can be used to characterize connections between users. Granovetter [8] advanced the weak-tie theory (i.e., distant and infrequent relationship) and proposed an efficient knowledge sharing by bridging the non-connected groups and individuals in an association's network.

Jin Xie et al. [9] proposed a recommendation method using collaborative filtering based on homophile and heterophile values. They considered user's behavior analysis in terms of similarity as well as diversity between their friends. This idea has a limitation in the scenario of dynamically changing behavior of followers in a Twitter-like social network. Table 1 summarizes the merits and demerits of various earlier work in this area.

These limitations of existing work motivate us for the design and development of recommendation system which can achieve more accuracy.

Table 1 Literature survey of existing works

Author	Merits	Demerits
Golder et al. [10]	Leveraged the idea of "Shared interest and Audience," "reciprocity," "Filterd people"	Did not validate their model
Gracia and Amatriain [11]	Identified important factors, e.g., Activity, popularity, locations, Tweet contents, etc. Good performance	Difficult to scale, not feasible to use for general public
Armentano et al. [12]	Explore the target user's followers and followees	Analysis performed just on a little number of users or nodes and further analysis is needed
Zhang L.Z., et al. [13]	Considered four interaction attributes (interaction frequency, interaction quality, seriousness, and common interest)	Limited to topology-based graphs and depends on user behavior in different environment
Shen D., et al. [14]	Modeling of user interest based on their blogs or their behavior history	Not feasible to all kind of users who do not have blogs or do not share behavior and interests

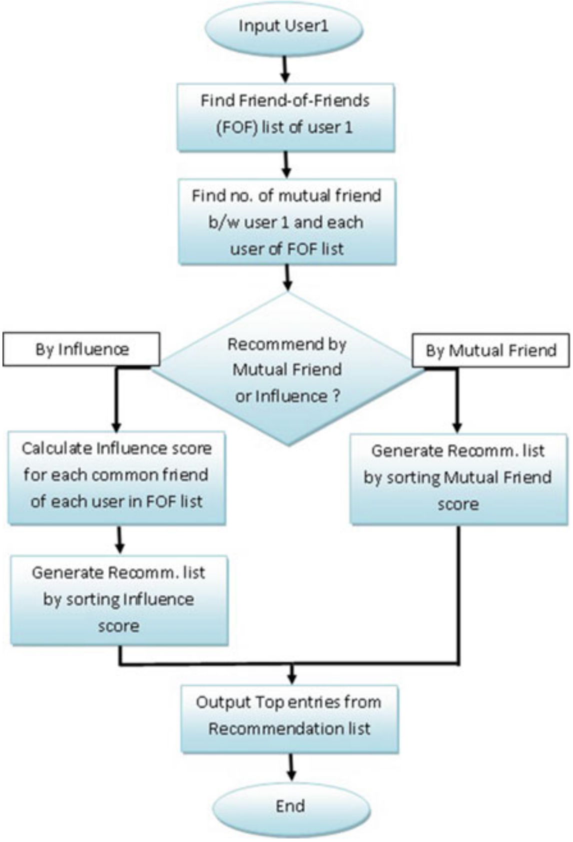
3 Proposed Methodology

In the context of development of recommendation system for online social networks or e-commerce sites following sections describe the proposed approach in details. Complete work flow of both proposed recommendation algorithms has been shown in Fig. 1.

3.1 Friendship Recommendation

In any social network, recommending a new friend is just the answer of a simple question "For user X, who is the best person to be recommended as new friend"? In order to get answer for this question for user X, we can make a list of some non-friends of X in order of best friend recommendation to worst friend recommendation.

Fig. 1 Work flow of proposed recommendation system


Input User1
Find Friend-of-Friends (FOF) list of user 1
Find no. of mutual friend b/w user 1 and each user of FOF list
By Influence
Recommend by Mutual Friend or Influence ?
By Mutual Friend
Calculate Influence score for each common friend of each user in FOF list
Generate Recomm. list by sorting Mutual Friend score
Generate Recomm. list by sorting Influence score
Output Top entries from Recommendation list
End

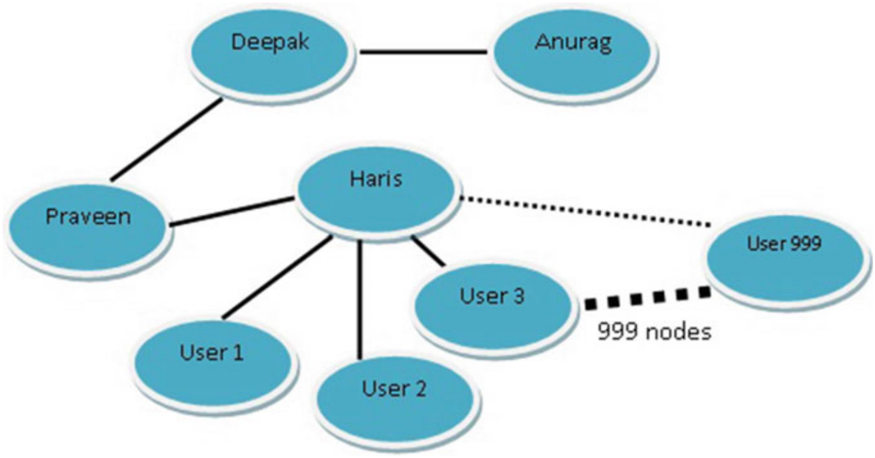


Fig. 2 Sample graph for demonstration of calculation of friendship and influence score

And we can select the top elements of list who is neither friend of X nor X itself. The recommended list may include all or a few of non-friends which depends on recommendation algorithm. Also a point here to be noted is that obtained result need not be symmetric always i.e. for user A, recommended friend can be B but for B it is not necessarily to be A.

Both algorithms proposed here, assigns a number called *score* to each potential friend. Higher is the score, higher is the similarity among users. Figure 2 shows the sample graph taken for demonstration of proposed algorithms.

3.2 Recommendation Based on Number of Mutual Friends

The idea here is that if X is friend of Y's friend then may be he should be Y's friend too. If many of Y's friends are friend of X then X may be an even better friend recommendation for Y. With whom the person Y is having maximum number of common friends is the best friend recommendation for Y. In the graph shown in Fig. 1 the *friendship score* of all these (non-friends with Praveen) users (Anurag, User1, User2, …, User999) is 1 because they have only one mutual friend with Praveen. So they all are equally likely to be recommended as a new friend for Praveen. The contribution of each common friend is 1 in the calculation of friendship score.

Algorithm 1 describes the procedure of the Mutual Friend based Recommendation system. There is a calculation of *friendship score* based on No of Common Friends between user and his friends-of-friends, which is a characteristic of the algorithm. It acts as the rank of user being recommended in the calculated recommendation list.

Algorithm 1: Recommendation based on Mutual Friends

```
Data: Graph dataset (G) having edge lists of connected users
Result: Getting Recommendation for friends
find Friends (G, user)
find Friends-of-friends (G, user, radius)
find Mutual_friends (G, user1, user2)
No_of_Mutual_friend_map(G, user)
    for each friend-of-friend
        calculate length_of (no. of common friend with user)
    return (friend-of-friend, rank)
Recommend_by_Mutual_Friend()
    Sort the list and return top entries from list as recommended friends
```

3.3 Recommendation Based on Influence

Recommendation based on the Influence uses the idea of prioritizing the nodes which are having more influence in recommendation. As a concrete example let us suppose a hypothetical situation for the graph shown in Fig. 2:

- Deepak and Haris are two of Praveen's friends.
- Deepak is having only two friends (Praveen and Anurag).
- Haris is having total 1000 friends (including Praveen).
- Deepak and Haris do not have any mutual friend.

Since Deepak is highly selective in making his friends and Praveen is one of his friend, so Praveen is likely to have a lot of common in Deepak's other friend Anurag. Whereas Haris is indiscriminate and there is not a strong reason to believe that Praveen should be friend with any particular friend of Haris. In this case the *Influence Score* of Anurag is 1/2, and each of Haris's friend would have a score of 1/1000. This can be understood as, each friend X of Praveen has a total influence score 1 to contribute in recommendation, and divides this influence score 1 equally among all friends of X.

Algorithm 2 describes the procedure of the Friendship Recommendation system based on the Influence. There is a calculation of *Influence* score based on idea that each friend of user has influence score of 1 which it divides among all of his friends, which is a characteristic of the algorithm. It also acts as the rank of user being recommended in the calculated recommendation list.

4 Results and Analysis

For Implementation purpose of our proposed algorithm, we took a sample graph consist of 11 nodes and made our algorithm generic for any large-scaled real Social Network Graph.

Algorithm 2: Recommendation based on Influence

```
Data: Graph dataset (G) having edge lists of connected users
Result: Getting Recommendation for friends
find Friends(G, user)
find Friends-of-friends(G, user, radius)
Influence_map(G, user)
    for each friend-of-friend
        find Mutual_friends(G, user1, user2)
        for each Mutual_friend
            influence_score=(1.0/len(friends(G, mutual_friend))
    return (friend, rank)
Recommend_by_Influence()
    sort the list and return top entries from list as recommended friends
```

Performance Metrics: For the measurement of degree of closeness of the obtained recommendation to the real preference of user, a numerical representation for accuracy has been considered. Accuracy is a widely used prediction metric in order to measure the quality of closeness of result obtained by an application with the real value, and for recommendation system it can be formulated as:

$$Accuracy = \frac{total\_no.\_of\_successful\_recommendations}{total\_no.\_of\_possible\_recommendations}$$

A recommendation can be considered successful if the recommended friendship is very close to the actual willing of user to make friendship with that recommended user.

Experimental Setup: In order to achieve and prove the effectiveness and efficient execution of our proposed algorithms we did experiment on large scale datasets of Facebook and Twitter available on http://www.snap.standford.edu [15] . These datasets are consist of 88,234 friendship edges between 4039 users and 1,768,149 friendship edges between 81,306 users of Facebook and Twitter respectively. We run our test on Quad core Intel i7-2600 with frequency 3.40 GHz and 8 GB RAM. Running these algorithms with these large data sets took around 20 and 75 s, respectively, on Facebook and Twitter data.

Experimental Result: We performed experiment and tested our proposed recommendation algorithm in the following steps:

1. Randomly selected a friendship connection of real friend from Graph data; say two friends User1 and User2.
2. From the graph we removed that friendship connection selected in step 1.
3. Calculated friendship recommendations for User1 and User2 using proposed algorithms.
4. Obtained the rank of User1 in the friend recommendation list of User2.
5. Obtained the rank of User2 in the friend recommendation list of User1.
6. Performed the accuracy measurement.

Table 2 Obtained recommendation list with score for user 1941 with randomly chosen 2000 edges 10 times for both recommendation algorithms

Trial no.	Recommendation by mutual friend method	Recommendation by influence method
1	[(‘2433’, 1), (‘2376’, 1), (‘2356’, 1)]	[(‘2026’, 0.33), (‘1947’, 0.33), (‘2433’, 0.16)]
2	[(‘2007’, 1), (‘1912’, 1)]	[(‘2007’, 0.5), (‘1912’, 0.5)]
3	[(‘2187’, 2), (‘2600’, 1), (‘2589’, 1)]	[(‘2054’, 0.5), (‘2187’, 0.41), (‘2600’, 0.25)]
4	[(‘2532’, 1), (‘2471’, 1), (‘2464’, 1)]	[(‘1997’, 0.5), (‘2471’, 0.33), (‘2328’, 0.33)]
5	[(‘2223’, 1), (‘2199’, 1), (‘2169’, 1)]	[(‘2223’, 0.5), (‘1948’, 0.5), (‘1912’, 0.5)]
6	[(‘2550’, 1), (‘2340’, 1), (‘2123’, 1)]	[(‘2032’, 0.5), (‘1912’, 0.5), (‘2550’, 0.25)]
7	[(‘2294’, 1), (‘2223’, 1), (‘2032’, 1)]	[(‘1940’, 0.5), (‘2294’, 0.33), (‘2223’, 0.33)]
8	[(‘2533’, 1), (‘2462’, 1), (‘2267’, 1)]	[(‘2533’, 0.5), (‘2267’, 0.5), (‘2203’, 0.33)]
9	[(‘2059’, 2), (‘1953’, 2), (‘2468’, 1)]	[(‘2059’, 0.53), (‘1953’, 0.53), (‘2071’, 0.5)]
10	[(‘2414’, 1), (‘2244’, 1), (‘2147’, 1)]	[(‘2147’, 0.33), (‘2003’, 0.33), (‘1945’, 0.33)]

Accuracy Measurement: For both recommendation algorithms proposed here, we performed the above experiment 10 times each by increasing the number of chosen friendship edges (10000, 20000, 30000, …, 80000, 88234) using the Facebook standard data. Within the recommendation list we computed the average rank of the correct recommendation. While computing the average rank that particular trial was ignored when recommendation list did not contain correct recommendation.

To prevent our results being skewed by different random choices, we evaluated both recommendation algorithms with the same random choice. In other Words it can be said that the random choices made each time, has been used to evaluate both recommendation systems, then went for the next choice.

Obtained results shown in Table 2 clearly indicate that there is a significant difference in recommendation list between proposed Algorithms 1 and 2. In the Table 3, we show the accuracy of both algorithms for different number of randomly selected edges.

Comparison graph of these algorithms shown in Fig. 3 clearly describes that recommendation system based on Influence gives recommendations with more accuracy and is very close to the real choice of the user to make new friends. Both algorithms perform well on large scaled data sets.

Table 3 Obtained accuracy with randomly chosen edges based on 100 trials for both recommendation algorithms

No of nodes taken randomly	Accuracy with mutual based recommendation	Accuracy with influenced based recommendation
10000	57.4	82.8
20000	63.6	86.1
30000	67.2	74.3
40000	74.7	89.4
50000	83.0	87.2
60000	85.4	86.1
70000	92.8	88.0
80000	90.5	96.5
88234	87.1	97.2

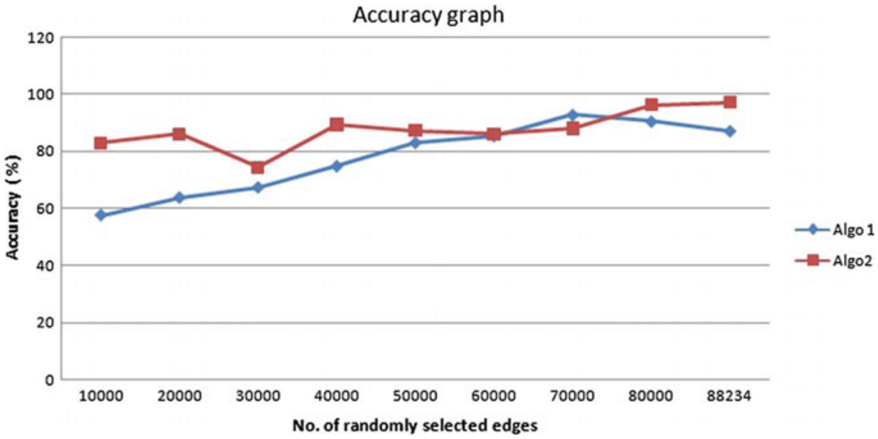


Fig. 3 Accuracy graph for proposed Algorithms 1 and 2

5 Conclusion and Future Work

Recommendation algorithms for Online social networks or e-commerce sites can be designed by considering several factors including behaviors, locations, similarity interest, features of products, etc. based on collaborative filtering or content-based filtering or a combination of both. Content-based filtering is lesser effective for recommendation in comparison of collaborative filtering as it is limited to recommending content of the same type as the user is already using. In this work two collaborative filtering-based recommendation algorithms have been proposed and their comparative study shows that recommendation based on influence gives more accurate recommendations as compared to common friend based recommendation. Processing of proposed algorithms with a very large scaled graph data sets of social networks is less efficient for centralized system with general configuration, as it is less suitable for dynamically increasing and scaling number of nodes day by day. This need lossless partitioning and distribution of nodes for efficient execution of these algorithms in distributed system, we consider it as our future work.

References

1. Linden, Greg, Brent Smith, and Jeremy York. “Amazon. com recommendations: Item-to-item collaborative filtering.” Internet Computing, IEEE 7.1 (2003): 76–80.
2. Sarwar, Badrul, et al. “Item-based collaborative filtering recommendation algorithms.” Proceedings of the 10th international conference on World Wide Web. ACM, 2001.
3. Van Meteren, Robin, and Maarten Van Someren. “Using content-based filtering for recommendation.” Proceedings of the Machine Learning in the New Information Age: MLnet/ECML2000 Workshop. 2000.
4. Jaiswal, A., S. Domanal, and G. R. M. Reddy. “Enhanced Framework for IoT Applications on Python Based Cloud Simulator (PCS).” 2015 IEEE International Conference on Cloud Computing in Emerging Markets (CCEM). IEEE, 2015.
5. Resnick, Paul, et al. “GroupLens: an open architecture for collaborative filtering of netnews.” Proceedings of the 1994 ACM conference on Computer supported cooperative work. ACM, 1994.
6. Kautz, Henry, Bart Selman, and Mehul Shah. “Referral Web: combining social networks and collaborative filtering.” Communications of the ACM 40.3 (1997): 63–65.
7. Garton, Laura, Caroline Haythornthwaite, and Barry Wellman. “Studying online social networks.” Journal of ComputerMediated Communication 3.1 (1997): 0–0.
8. Granovetter, Mark S. “The strength of weak ties.” American journal of sociology (1973): 1360–1380.
9. Xie, Jin, and Xing Li. “Make best use of social networks via more valuable friend recommendations.” Consumer Electronics, Communications and Networks (CECNet), 2012 2nd International Conference on. IEEE, 2012.
10. Golder, Scott A., et al. “A structural approach to contact recommendations in online social networks.” Workshop on Search in Social Media, SSM. 2009.
11. Garcia, Ruth, and Xavier Amatriain. “Weighted content based methods for recommending connections in online social networks.” Workshop on Recommender Systems and the Social Web. 2010.
12. Armentano, Marcelo G., Daniela L. Godoy, and Anala A. Amandi. “A topology-based approach for followees recommendation in Twitter.” Workshop chairs. 2011.
13. Zhang, Lizi, et al. “IntRank: Interaction ranking-based trustworthy friend recommendation.” Trust, Security and Privacy in Computing and Communications (TrustCom), 2011 IEEE 10th International Conference on. IEEE, 2011.
14. Shen, Dou, et al. “Latent friend mining from blog data.” Data Mining, 2006. ICDM’06. Sixth International Conference on. IEEE, 2006.
15. Dataset taken from Link: http://snap.stanford.edu/data/index.html#socnets

Personalized Recommendation Approach for Academic Literature Using High-Utility Itemset Mining Technique

Mahak Dhanda and Vijay Verma

Abstract As the size of digital academic library is increasing enormously, it has become arduous for the researchers to identify the papers of their interest from this repository. This has escalated researcher's attention toward the implementation of Recommender Systems (RS) in academic literature domain. The content-based and collaborative filtering-based techniques when applied in the academic literature domain, fail in reflecting the researcher's personalized preferences in terms of recentness, popularity, etc. This article presents a Personalized Recommendation Approach for Academic Literature which is based on High-Utility Itemset Mining (HUIM) Technique. This approach uses the content of the paper along with user's personalized preference, for making recommendations. Here, we have utilized a highly efficient HUIM algorithm, EFIM, which has been recently introduced in the literature, to mine the papers having higher utility to the user. Experimental evaluation proves that our work satisfies the researcher's personalized requirements and also outperforms the existing personalized research paper recommender systems in terms of its time and space complexities.

Keywords Academic literature · Content-based filtering · Minimum utility threshold · High-utility itemset mining · Recommender system

1 Introduction

The exponential growth of information on the web has made it onerous for the users to find the information pertinent to them. RS has emerged as revolutionary idea to endure this situation. It is a software tool that could suggest the user with the

M. Dhanda (✉) · V. Verma
Department of Computer Engineering, National Institute of Technology Kurukshetra,
Haryana 136119, India
e-mail: mahak0570@gmail.com

V. Verma
e-mail: vermavijay1986@gmail.com

P.K. Sa et al. (eds.), *Progress in Intelligent Computing Techniques: Theory, Practice, and Applications*, Advances in Intelligent Systems
and Computing 719, DOI 10.1007/978-981-10-3376-6_27

information to be of their use. Many recommendation approaches are available presently that can me majorly classified in two classes: collaborative filtering and content-based filtering. In collaborative filtering recommendation is made on the basis of ratings by the similar users, whereas in content-based techniques recommendation is made by finding the similarity between item's features and features of user as per his profile.

RSs can be adapted to work in academic literature domain also. As the size of the academic literature library is growing enormously, researchers find it burdensome task to find the research papers akin to them. To get by this situation, we here propose a personalized recommender system for research papers that takes into consideration user's personalized requirements along with the content quality. In this, the HUIM technique is used to find the degree of relevance of a research paper to the user and recommend only the High-Utility Reference-sets (HURs). The proposed technique works in two steps (1) research papers akin to user's topic of interest are selected using content-based filtering method; (2) from the selected research papers HURs are selected (whose utility cross the minimum utility threshold ϴ provided by user).

The rest of the article is organized as: Sect. 2 summarizes the related development in the concerned field; Sect. 3 details the proposed approach; Sect. 4 contains the simulation results and conclusion is added as Sect. 5.

2 Related Work

RSs are responsible for suggesting the users with information that may be pertinent to them. The recommendation approaches are mainly classified in two groups (1) collaborative filtering; (2) content-based filtering. Both types of approaches focus on relevancy of information to the user on the basis of either user's analogous users or user's own profile without considering any customization of preferences [1–3].

RSs have become an important part in many application domains, e.g., e-commerce [4], e-learning [5], etc. Research paper Recommender Systems tend to recommend papers either on the basis of its citation frequency or by the relevance of its contents to the user but do not contemplate user's personalized preferences [6, 7]. Recommendation approach based on utility has been developed to address user's personalized requirements but lacks in efficiency because it relies upon Two-phase algorithm [8], for mining HURs, which has large space and time requirements [9].

3 Proposed System

Our developed approach is a personalized recommendation approach for academic literature that works using HUIM technique EFIM. The entire task of recommendation is accomplished in two steps (1) selecting the papers akin to user's topic of interest; (2) recommending papers having higher usefulness to the user.

3.1 *Architecture of the System*

The basic architecture of the system is as shown in Fig. 1.

3.2 *Working*

Step I **Filtering on the basis of research paper's content**:

In this step the complete repository of research papers is partitioned into 'k' clusters using the PLSA algorithm. Each cluster is accredited with a topic implicitly and then for each paper its probability of fitting to a particular cluster is calculated on the basis of their word distributions. A paper is assigned to the cluster with which its association probability is greatest. Once the clusters are made closeness of each cluster to user's topic of interest is determined using the similarity measures. The cluster with highest affinity toward user's topic of interest is selected to work upon further. While all the other clusters are discarded avoiding unnecessary overhead.

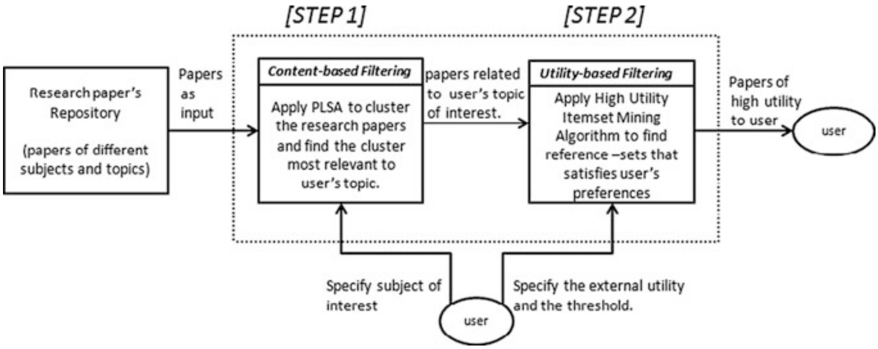

[STEP 1]
[STEP 2]
Research paper's Repository
(papers of different subjects and topics)
Papers as input
Content-based Filtering
Apply PLSA to cluster the research papers and find the cluster most relevant to user's topic.
papers related to user's topic of interest.
Utility-based Filtering
Apply High Utility Itemset Mining Algorithm to find reference –sets that satisfies user's preferences
Papers of high utility to user
user
Specify subject of interest
Specify the external utility and the threshold.
user

Fig. 1 Architecture of the system

Step II **Filtering on the basis of user's personalized requirements**:

In order to meet user's personalized requirements (authority, publishing date, etc.) the system employs the HUIM technique, EFIM, which will provide the user with the reference-sets that best suits his requirements. HUIM algorithms are used to find out the items with utility crossing Θ. In academic domain, utility specifies degree to which paper is preferred by the user. It can be specified on the basis of date of publishing or publishing authority, etc. In this domain, EFIM is applied on papers instead of the transactions. Also items are replaced by references and itemsets with reference-sets. In this, the *internal utility* (*i*) of a reference is either 1 or 0 defining if a reference is used in a paper or not. The *external utility* (*e*) is provided by user according to his preferences (which is publishing date here). The *utility of a reference in any paper*, u(r, P) = i * e where i and e are the internal and external utilities of a reference, respectively, and the *utility of a reference-set* (*R*) is the sum of utilities of all the references present in the set. A utility threshold Θ will be provided by the user which defines the minimum utility bound to be satisfied by the paper to get recommended. Only those reference-sets are then recommended whose utility ≥ Θ and for catching out these reference-sets HUIM algorithm EFIM is used. EFIM discovers all the HURs in a single phase. For this it uses the concept of projected database, transaction merging, remaining utility, and utility list [10]. From the set of HURs found out by EFIM top most HURs are recommended to user.

3.3 Example

Consider the dataset in Table 1 as running example. Here, columns represent the references used in papers while rows represent the papers present in the dataset. Entry in each cell denotes the number of times citation is given to a reference in a paper. Table 2 gives the utility value (weights) assigned to each reference by the user.

Table 1 The research paper repository

ID of paper	Reference r1	Reference r2	Reference r3	Reference r4
P1	1	1	0	1
P2	0	0	1	0
P3	0	0	1	0
P4	0	1	0	0
P5	1	0	1	1

Table 2 Weightage given to references by the user

Reference	r1	r2	r3	r4
Weight Assigned	4	3	1	1

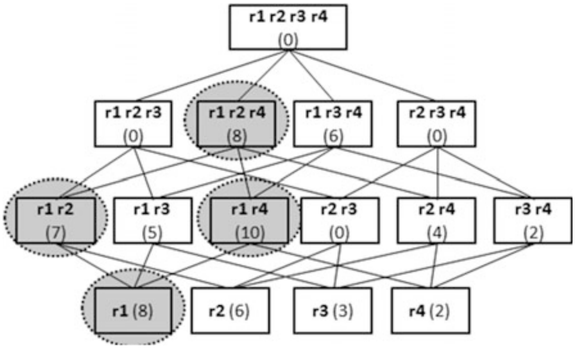


Fig. 2 Possible reference-sets

Here utilities of different reference-sets are: u({r1}, P1) = 4, u({r1}) = 8, and u ({r1, r4}) = u({r1, r4},P1) + u({r1, r4}, P5) = 5 + 5 = 10.

If the Θ is set at 9, reference-set {r1, r2} will not be recommended. But if Θ is 7, {r1, r2} will be a HUR and hence recommended. If we consider recommendation based on citation frequency and set Θ as 3 then r3 will be recommended to the user but if we check the utility-based recommendations r4 is not HUR as u({r3}) = 3 which is less than the Θ, i.e., 7. Hence r3 is not recommended to the user which indicates that r3 cannot fulfill user's personalized requirements. All the possible reference-sets for the above example are shown in Fig. 2. The numeric values denote the utility value of that reference set. If Θ is set as 7 then only the reference-sets with utility value ≥ Θ will be HURs and hence will be recommended to the user. Reference-sets {r1}, {r1 r2}, {r1 r4}, {r1 r2 r4} are HURs found out by EFIM (in dark solid circles) and hence are recommended to the user.

4 Simulation Results

4.1 Data Set

To evaluate our proposed approach, a real-world dataset, ACL Anthology Network [11] is used. It is a citation network of research papers collected by ACL from its various venues. Because the topic of data set is already consistent, we will directly mine HURs from the database. There is no need to perform filtering on the basis of contents. We are considering the collection of research papers with publishing date between 1965 and 2007. The "publishing date" is used as the factor through which personalized requirements of the user will be fulfilled. The weightage will be provided to the research papers explicitly by the user on the basis of their publishing date. Preprocessing of the data is done. The ACL repository contains various research papers along with the citation network of those papers. To apply HUIM techniques each paper Pi is treated as a transaction with references denoting the items in transaction, to make transaction-itemset network. For this some IDs (unique) have to be assigned to the papers and papers are represented in the results as these IDs only.

4.2 Results

We have performed various experiments to check the capability of our system as:

1. Recommendation on the basis of citation frequency, where the weight of papers on the basis of their publishing date is same.
2. Recommendation on the basis of utility of reference-set using Two-phase algorithm. The references with publishing date as 2005–2007, 2000–2004, 1990–1999, 1965–1989, get the weightage 20, 10, 5, and 1, respectively.
3. Recommendation on the basis of utility of reference-set using EFIM algorithm. The references with publishing date as 2005–2007, 2000–2004, 1990–1999, 1965–1989, get the weightage 20, 10, 5, and 1, respectively.

In experiment 2 and 3 we have used ϴ as 2000. ϴ can be given any random value depending on the requirements of the user.

In Table 3, the variations between the results of 1 and 2 experiments are shown as bold entries. The reference-sets {2480, 1953}, {2598, 1953}, {2480, 1869}, {2598, 1950} are discovered by the EFIM-based filtering but not by the filtering on the basis of citation frequency. Whereas the reference-sets {1869, 13}, {1950, 13}, {1902, 13} are discovered by the citation frequency-based filtering but not by EFIM-based filtering. The references 2480 and 2598 are published in year 2007 and 2005, respectively, which have been provided with higher weightage by the user (because of their recentness). Whereas the reference 13 is published in year 1993 so less weightage is given to it by the user (as it is older). Hence the reference-sets including 13 are not HURs. This proves that our proposed approach not only brings the HURs into the recommendation list but also removes the reference-sets with

Table 3 The reference-sets obtained after performing experiment 1, 2, and 3

Rank	Experiment 1	Experiment 2	Experiment 3
1.	1953 1902	1953 1902	1953 1902
2.	1950 1953	1950 1953	1950 1953
3.	**1869 13**	**2480 1953**	2480 1953
4.	1950 1902	**2598 1953**	2598 1953
5.	1950 1953 1902	**2480 1869**	2480 1869
6.	1869 1950	1950 1902	1950 1902
7.	1869 1902	**2598 1950**	2598 1950
8.	**1950 13**	1950 1953 1902	1950 1953 1902
9.	1869 1953	1869 1950	1869 1950
10.	**1902 13**	1869 1902	1869 1902

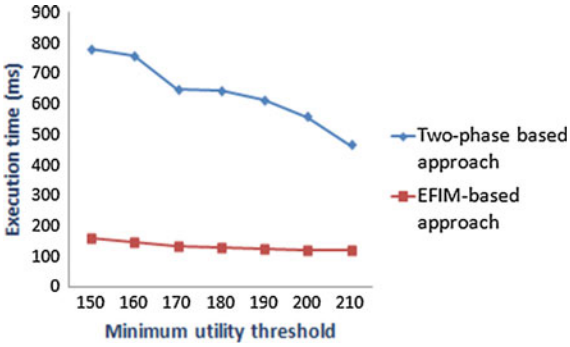


Fig. 3 Comparison of execution time

lesser utility. As the reference-sets appearing in columns third and fourth are same, this proves that both the utility-based recommendation approaches (Two-phase-based and EFIM-based) provides same results but the difference lies in the time and space complexities of both.

The EFIM-based recommendation approach consumes less space and time than Two-phase because no intermediate results are generated (no candidate-generation step) and also it uses highly efficient search space pruning techniques. Graph in Fig. 3 verifies that the proposed EFIM-based approach is four to five times faster than the Two-phase based approach for every minimum utility threshold.

5 Conclusion

In this article, we have proposed a Personalized Recommendation Approach for Academic literature using High-Utility Itemset mining Technique. We have used the concept of utility for this that helped to provide recommendations to the user by taking into account their personalized requirements also. The utility of a reference is defined in two parts as internal and external utility. The approach not only provides quality in recommendation by filtering on the basis of contents but also satisfies user's personalized requirements by utility-based filtering. EFIM algorithm is followed to give high-utility reference-sets as output. Simulation results prove that different reference-sets are recommended changing the external utility value and also the proposed approach successfully fulfills the user's personalized requirements.

References

1. de Gemmis, Marco, et al.: Semantics-Aware Content-Based Recommender Systems. Recommender Systems Handbook, 119–159 (2015).
2. Koren Y., Bell R.: Advances in collaborative filtering. Recommender Systems Handbook, 77–11 8(2015).
3. Aggarwal, Charu C.: Ensemble-Based and Hybrid Recommender Systems. Recommender Systems. Springer International Publishing, 199–224 (2016).
4. Boehmer J., Jung Y., and Wash R.: e-Commerce Recommender Systems. The International Encyclopedia of Digital Communication and Society (2015).
5. De Maio C., et al.: Rss-based e-learning recommendations exploiting fuzzy for knowledge modeling. Appl. Soft Computing, 113–124 (2012).
6. Champiri Z D., et al.: A systematic review of scholar context-aware recommender systems.: Expert Systems with Applications, Elsevier, 1743–1758 (2015).
7. Beel J., et al.: Research-paper recommender systems: a literature survey. International Journal on Digital Libraries, 1–34 (2015).
8. Liu Y., et al.: A fast high utility itemsets mining algorithm. Proceedings of the 1st international workshop on Utility-based data mining, ACM, 90–99 (2005).
9. Liang S., et al.: A Utility-based Recommendation Approach for Academic Literatures. ACM International Conferences on Web Intelligence and Intelligent Agent Technology, 229–232 (2011).
10. Zida S., Fournier-Viger P., Lin J.C.-W.: EFIM: A Highly Efficient Algorithm for High-Utility Itemset Mining. Proc. 14th Mexican Intern. Conference on Artificial Intelligence, Springer, 530–546 (2015).
11. ACL Anthology Network, http://clair.si.umich.edu/clair/anthology/.

An Estimation of User Preferences for Search Engine Results and its Usage Patterns

Nidhi Bajpai and Deepak Arora

Abstract The fields where an exhaustive understanding of user preferences can be applied include web page ranking, web search personalization, and search engine adaptation. The most important use of an understanding of how people use search engines and what they want from them is the immense scope it creates for system improvement. System improvement means evolving search engines to constantly exceed user expectations. Various approaches like creating user profiles, saving logs of user search patterns, evaluating users' browsing behavior, etc., have been used to determine user preferences. This research work aims to determine different aspect of user preferences in respect of Search Engines. The authors present a method to analyze and evaluate user preferences for search engines based on an experiment, which was conducted on working professionals employed in various domains like software companies, law firms, banks, educational institutes, government, etc. The sample of the study has 120 working professionals working in different domains.

Keywords Search engine · User preference · Internet · Search engine result

1 Introduction

Search engines have seen revolutionary changes in both the way they work as well as the users' perspective of using them. A study of user preferences is required to have a greater understanding of the needs and expectations of users from search engines. The relevant results from such a study can be used in development, deployment, and maintenance of search engines [1].

N. Bajpai (✉) · D. Arora
Department of Computer Science & Engineering, Amity School of Engineering,
Amity University, Lucknow, India
e-mail: nidhibajpai07@gmail.com

D. Arora
e-mail: deepakarorainbox@gmail.com

P.K. Sa et al. (eds.), *Progress in Intelligent Computing Techniques: Theory, Practice, and Applications*, Advances in Intelligent Systems and Computing 719, DOI 10.1007/978-981-10-3376-6_28

This research work intends to determine and evaluate user preference for search results and usage, which will help in taking future steps for system improvement. The authors have presented different aspects of user preferences. These aspects are studied and analyzed from users' perspective and domain wise analysis of theses aspects is performed. Different domains include software companies, law firms, banks, education, government offices, and 'others'.

This research work starts with the detailed study of aspects of user preferences for search engine results and usage with users' response on each aspect. After that, the methodology used in the research work is presented. This is followed by evaluation, analysis, and graphical representation of results and domain wise analysis of these results. Finally, the conclusion and future scope is defined.

2 Identification of Different Aspects of User Preferences and Experimental Setup

In order to derive these preferences, the authors conducted an online experiment on working professionals employed in different domains as indicated above. In this experiment, few questions related to search engine were presented to the participants. Based on the input received from the participants, the authors have determined different aspects of user preferences. These different aspects of user preference are discussed in the headings below.

2.1 User Preference for Private Search

When users are searching for information using search engine, different logs are created both at the search engine side and browser side [2]. Search engine logs contains details like query string, timestamp of query, IP address, click through URLs, search results,and links traversed by the user [2]. This information is logged for various purposes. Thus, it is evident that in normal search engine browsing, large amount of details related to the user are logged or saved in the system. Private search provides freedom to its user to search without the aforesaid details being logged. Private search disables all the logs and cookies, and no information is saved in regards to the user and his/her search details, etc. Private search aims to respect users' privacy while searching information through search engines. Private search is thus one of the aspects of user preference. The findings to this aspect are shown in Table 1. Show three important results:

- It was found that 65% of working professionals in different domains feel that privacy is an issue while using search engines.
- The majority of working professionals surveyed, i.e., 87.50% of working professionals prefer private search over normal search.

Table 1 Findings for private search

	Value	Frequency	Percent (%)
Do you feel privacy issue while using search engine?	Yes	78	65
	No	42	35
Do you want your search history to be private?	Yes	105	87.50
	No	15	12.50
How much are you willing to pay for private search?	I will not pay	98	81.67
	Up to Rs. 500 per month	18	15
	Up to Rs.1000 per month	4	3.33

Table 2 Findings for preference for number of search results

	Frequency	Percent	Valid percent	Cumulative percent
Less but accurate results	73	60.8	60.8	60.8
More no. of results	16	13.3	13.3	74.2
Optimum result	31	25.8	25.8	100.0
Total	120	100.0	100.0	

- Out of 120 professionals, 98 professionals are not willing to pay at all for the private search; however, 22 said that they were willing to pay for private search.
- This shows that 18.33% of working professionals are even willing to pay for private search.

2.2 *Preference for Number of Search Results*

In order to retrieve information from search engines, users input queries in the given query box. In response to this user query, search engines display a result page which consists of different results arranged according to the ranking algorithm used by a particular search engine [3]. A search engine result page (SERP) consist of a number of results and search results are divided over various SERP based on the search engine used.

Findings on this aspect of user preference are shown in Table 2. It is said that 86.67% of working professionals are interested in less number but optimum results, while only 13.33% of users were interested in large number of results and SERPs. This clearly indicates that there is need to optimize the accuracy of results shown in the SERP rather than showing large number results in a SERP or showing large number of SERP.

2.3 Generalized Results Versus Customized Results Versus Personalized Results

Personalized results are created by a filter that takes various parameters like users' browsing history, past browsing preferences and location into account [4]. Customized results were also one option in this experiment. Customized results in this experiment are described as results, which are based not only on the query but also the custom options selected by the user at the time of submitting a query. Generalized results are results,which are displayed based on ranking algorithms only and no other filtering of user options or history is taken into account.

The findings for this aspect of user preference are shown in Table 3. It is said that majority of users (54.17%) are interested in customized results, which gives more freedom to user in terms of filtering options and gives a more focused view of the web as compared to generalized search. Second preference among users is for generalized results (32.50%), which gives complete view of the web to the user. The least preferred search results are for personalized results with only (13.33%) of the respondents' opting for them.

There is a common criticism that personalized search does not provide a complete view of the web to the user [5]. There was much hype with the release of personalized search but according to this research work, personalized results are least preferable as per the statistics.

2.4 Organic Versus Sponsored Results

Organic results are the 'natural results', which are based on the relevance of the user search query and are ranked accordingly whereas sponsored results are paid advertisements [6]. Meaning that the web page owner has to pay to have their link displayed on SERP for certain keywords. Payments are made on the basis of clicks made on the link for the advertisement this is known as pay per click (PPC).

The users can easily distinguish between organic result and sponsored result because the search engine keep them separate by placing them to the right of organic result or sometimes they are placed above the organic results, sometimes sponsored links have shaded background, border lines and other formatting features which keeps them separate from organic results.

Table 3 findings for type of results

	Frequency	Percent	Valid Percent	Cumulative Percent
Customized search	65	54.2	54.2	54.2
Generalized results	39	32.5	32.5	86.7
Personalized results	16	13.3	13.3	100.0
Total	120	100.0	100.0	

Table 4 Do you click on advertisement which appear while using search engine?

	Frequency	Percent	Valid percent	Cumulative percent
No	103	85.8	85.8	85.8
Yes	17	14.2	14.2	100.0
Total	120	100.0	100.0	

Table 5 Findings for number of search engine result page

	Frequency	Percent	Valid percent	Cumulative percent
More no of pages	33	27.5	27.5	27.5
Page 1 results	47	39.2	39.2	66.7
Page 2 results	30	25.0	25.0	91.7
Page 3 results	10	8.3	8.3	100.0
Total	120	100.0	100.0	

The finding to this aspect is shown in Table 4. Says that 85.83% professionals never click on the sponsored link. In this way, the organic results that are the natural results are preferred by the users.

2.5 *Number of Search Engine Result Page*

A number of search engine result page (SERP) are displayed in response to every query entered by the user. The SERP are arranged based on the particular search engine-ranking algorithm. In this study, the authors asked the participants in the experiment about the number of SERP pages up to which their search is limited to.

The finding about this aspect is shown in Table 5. Says that 72.50% of users' search is limited from 1 to 3 pages only. Authors believe that more emphasis should be paid to first three SERP as majority of users search first three SERP for results.

2.6 *Search Bias Versus Search Neutrality*

Search neutrality is a principle that states the search results displayed by the search engine should be based only on the relevance and ranked according to them [7, 8]. The search results should be impartial and unbiased, whereas search bias is the opposite of search neutrality where the search engine favors few websites over others in the ranking order in the SERP displayed [7, 8]. Search neutrality favors organic results while search bias favors paid results and sponsored links. The users' opinion about search bias and search neutrality is analyzed in this research work.

According to this experiment, 48.33% feels that there is search neutrality while 51.67% feel that Search bias is there.

3 Methodology Used

3.1 Stage I—Qualitative Research

In this study, first the authors have used qualitative research methodology by doing detailed study about search engines. The exploratory research was done aimed to achieve better understanding of the topic. Literature review technique was used to gain better understanding of the topic, which helped further to frame research questions. This exploratory research step will aid in more powerful analysis of the subject at later stages. This step is important to have better understanding of the topic, users' perspectives and future implications of the topic and provides better insight into the topic.

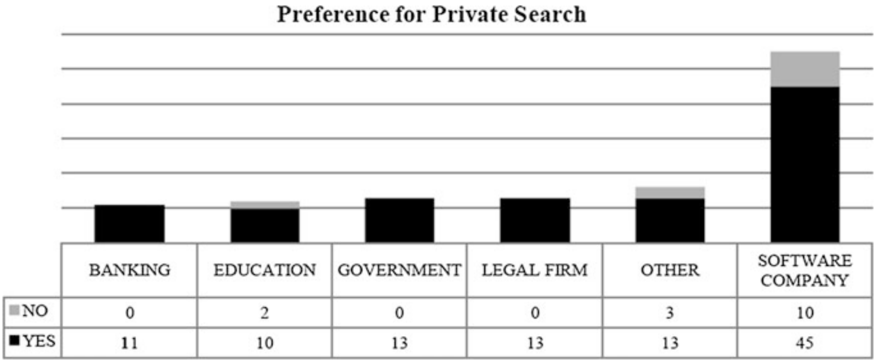
3.2 Stage II—Quantitative Research

A Quantitative research was followed which used the online survey methodology. A questionnaire was designed using Google forms and circulated online to working professionals in different domains like software company, government, banking, legal firms, education, etc. The responses collected from the working professionals were analyzed using the software GNU PSPP software.

4 Results and Discussions

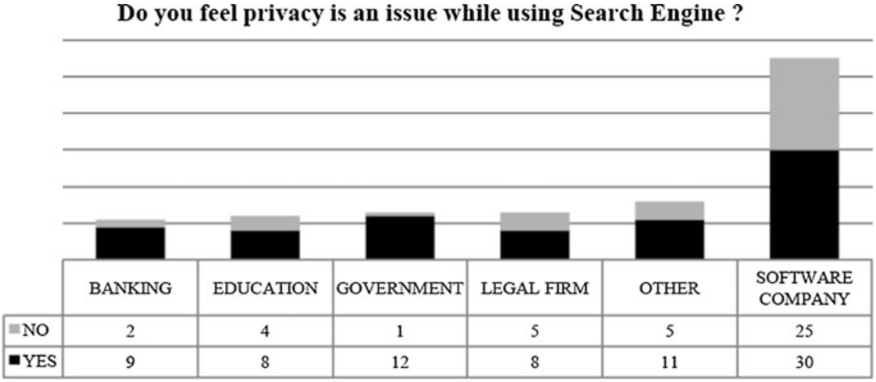
This research work is based on an experiment conducted online amongst working professionals. The sample consisted of 120 working professionals from different domains, which includes 55 professionals working in different software companies, 13 law professionals practicing independently or in legal firms, 13 professionals working in government services, 12 professionals working in education field, 11 professionals working in different banks, and 16 professionals in 'other' domains. The sample includes 76 male professionals and 44 female professionals.

The results about preference for private search are displayed in Fig. 1. Below which shows that the professionals in banking, government, and legal firms have 100% preference for private search. The authors explain this point with the fact that the task performed in banking, law firms and government is very critical and



	BANKING	EDUCATION	GOVERNMENT	LEGAL FIRM	OTHER	SOFTWARE COMPANY
▨ NO	0	2	0	0	3	10
■ YES	11	10	13	13	13	45

Fig. 1 Result for preference for private search



	BANKING	EDUCATION	GOVERNMENT	LEGAL FIRM	OTHER	SOFTWARE COMPANY
▨ NO	2	4	1	5	5	25
■ YES	9	8	12	8	11	30

Fig. 2 Result for privacy issue

confidential so professionals working in these domains have 100% preference for private search.

On the other hand, professionals in software company, education, and ‘other’ domains prefer private search but do not abstain from using normal search as well. While in case of participants from banking sector, government, and law firms there is an absolute preference for private search.

The result for domain wise analysis of privacy issue is shown in Fig. 2, which shows that in all the domains, professionals perceive a primacy regarding privacy issues while using search engine.

The result for preference of number of search result is shown in Fig. 3, which states that users have more preference for accurate but less number of results. The results show that 86.67% of users have preference for fewer but accurate results or optimum results while only 13.33% of users want greater number of results.

The result for domain wise analysis of organic versus sponsored links is shown in Fig. 4. Which shows that in banking domain, there is 100% preference for

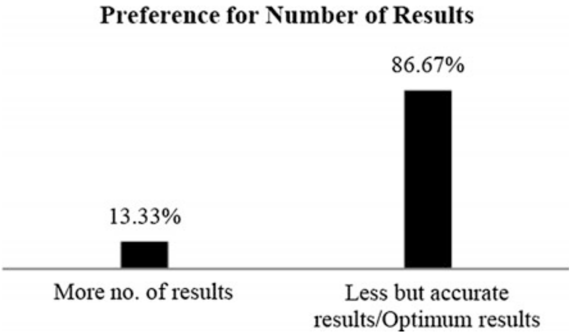


Fig. 3 Results for number of results

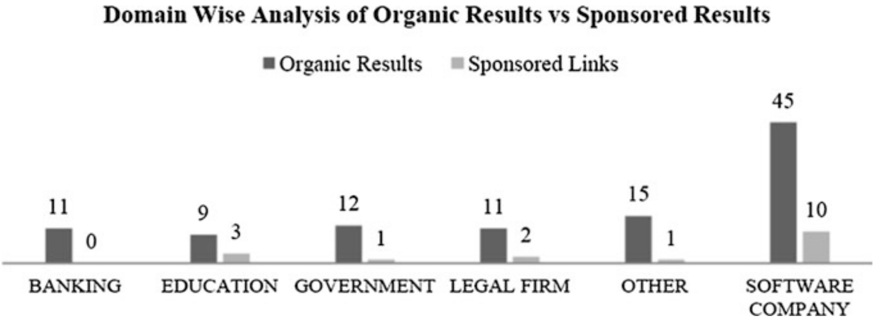


Fig. 4 Domain wise analysis of organic versus sponsored results

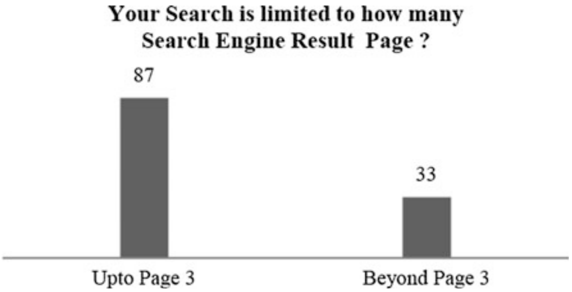


Fig. 5 Result for number of search results

organic results and no preference for sponsored links while in other domain the higher preference is for organic results and much lesser preference for sponsored links.

The result for number of SERP is shown in Fig. 5, which shows that 87% of users visit only first three SERP and only 13% go beyond first three SERP.

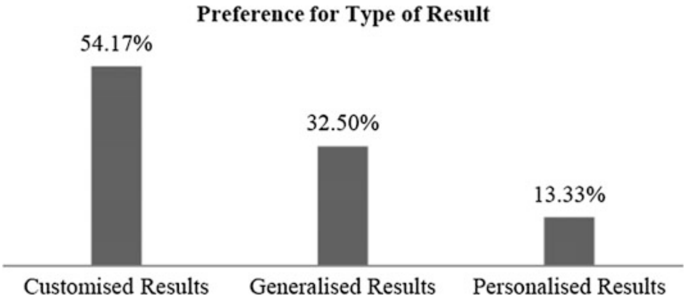


Fig. 6 Preference for type of result

From results shown in Figs. 3 and 5, the authors emphasize that there is need for fewer but accurate, specific, and relevant results rather than large number of search results.

The authors introduced the concept of customized results defined as results, which are based on the query and customization options selected by the user at that time of query. Customized results provide an option to users to select the filtering option at the time of query. In this way, customized results give more control to users' vis-a-vis results that the user is expecting from the search engine.

The findings of this experiment in respect of the type of search option are shown in Fig. 6. Below which states that users have highest preference for customized results, i.e., 54.17%; however, 32.50% users prefer generalized results while 13.33% users only prefer personalized results. Customized search that is preferred by majority of users is the combination of the best features of both generalized search and personalized search. Meaning thereby that it gives complete view of the web like Generalized results and combines filter based on users' choice like personalized results but the option of filtering is provided at the at the time of query, thus giving a greater room for customization of results.

5 Conclusion

In this study, authors have presented analysis of users' preferences for search results and usage in different domains based on an experiment performed across a cross section of working professionals. As shown in results, majority of users prefer private search with a finite amount of accurate and specific organic search results. The users prefer customized search results more with generalized search results as the second preference. So in that way, customized research is more demanding area, in which most of the browsers needs to improve. This analysis also shows that more emphasis should be paid to results shown in first three SERP, as users prefer to have their results in first three SERP.

This research work also introduced the concept of customized search result, and also analyzed the user preference about search neutrality and search bias concerns related to search engines. This research work also focuses on the search result accuracy of a browser, which will depend user to user, what they want to search or expect. The results of this experiment will be very helpful for the upcoming researchers towards increasing the efficiency and opening new dimensions of different search algorithm/techniques, by adding more intelligence to upcoming browsers in the market. The future scope of this study includes conducting the same experiment on larger sample size by including more different domains for further analysis and domain specific understanding.

References

1. Eugene, Agichtein et al. "Learning User Interaction Models For Predicting Web Search Result Preferences". Proceeding SIGIR '06 Proceedings Of The 29Th Annual International ACM SIGIR Conference On Research And Development In Information Retrieval. New York: ACM, 2006. 3–10.
2. Jiang, Daxin et al. "Mining Browse And Log Search For A Web Search". ACM Transactions on Intelligent Systems and Technology 4.4 (2013).
3. "What Is Google Pagerank? A Guide For Searchers & Webmasters. Search Engine Land". www.searchengineland.com. Web.
4. James, Pitokow et al. "Personalised Search". Communications of the ACM 45.9 (2002): 50–55.
5. Pariser, Eli. The Filter Bubble: What The Internet Is Hiding From You By. Penguin Group, 2011.
6. Jansen, Bernard, and Marc Resnick. "An Examination Of Searcher's Perceptions Of Non-Sponsored And Sponsored Links During Ecommerce Web Searching". Journal of the American Society for Information Science and Technology 57.14 (2006): 1949–1961.
7. Raff, Adam. "Search, But You May Not Find". The New York Times 2009.
8. Lao, Marina. 'Neutral' Search As A Basis for Antitrust Action?. Harvard Journal of Law & Technology Occasional Paper Series. 2013 July:1–12.

A Comparative Analysis of Various Spam Classifications

Nasir Fareed Shah and Pramod Kumar

Abstract Bandwidth, time, and storage space are the major three assets in computational world. Spam emails affect all the three, thus degrade the overall efficiency of the system. Spammers are using new tricks and traps to land these frivolous mails into our inbox. To make mailboxes more intelligent, our effort will be to devise a new algorithm that will help to classify emails in much smarter and efficient way. This paper analyzes various spam classification techniques and thereby put forward a new way of classifying spam emails. This paper thoroughly compares the results that various authors have got while simulating their architectures. Our approach of classification works efficiently and more accurately on varied length and type of datasets during training and testing phases. We tried to minimize the error ratio and increase classifier efficiency by implementing Genetic Algorithm concept.

Keywords Spam classification · Spam email · Unsolicited · Feature set · Logistic regression · Genetic algorithm · Machine learning

1 Introduction

Unsolicited bulk email or junk email are frivolous mails, which are sent in bulk to either make an advertisement [1], proliferate viruses, hack mailboxes [1], cheat somebody, or send a prank. As emails are sent to millions with no incurring cost, the spam traffic between MTA's causes delayed delivery of true mails [2]. Spams nearly occupy about two-third of our mailboxes [1], thereby causing inefficient utilization of storage space, bandwidth, and time [1].

N.F. Shah (✉) · P. Kumar
Department of Computer Science & Engineering, Birla Institute of Technology,
Mesra, Ranchi 835215, India
e-mail: saednasir@gmail.com

P. Kumar
e-mail: mepramod19@gmail.com

P.K. Sa et al. (eds.), *Progress in Intelligent Computing Techniques: Theory, Practice, and Applications*, Advances in Intelligent Systems
and Computing 719, DOI 10.1007/978-981-10-3376-6_29

In order to keep spammers at bay, there are many spam filtering techniques which are robust enough to detect a spam mail. Some of them use knowledge Engg. (KE) based approach, while majority of them are following the machine learning (ML) approach [3]. The latter is more robust and intelligent way of classifying emails. The former uses the stored procedure or rules to classify emails. It may have stored dictionary of words like BUY, SPAM, Lottery, Offer, Prize, Reward, etc. It periodically updates its dictionary to adapt with new trends [4]. But this practice is not so efficient because once the dictionary or repository of words is set, it is impossible to constantly update it at different end-user sites.

In comparison to KE, machine-learning approach (ML) is an intelligent way of filtering spams. ML do not have predefined rules or procedure. It can mutate itself to adapt with user needs, so ML is based on user adaptability. Our research will be based on the analytical approaches put forth by various researchers. We will thoroughly analyze their approaches and results, thereby devise a new spam classifying tool and deduce efficient results.

2 Various Spam Filtering Classification Techniques

2.1 Classification Based on Advanced Feature Abundance [1]

In this approach, the authors take raw emails as input, and label each email as fraud or normal. There architecture has following components:

Input module: In this section, the raw emails are fed from the external source (database repository) into the filter. The raw emails are supposed to contain two parts, (a) **Header** and (b) **Body**.

Content extractor: Here the two vital contents (subject and Body) of raw email are extracted. The subject parts are more indicative of action or process which the spammer is intended to do. The body part contains the detailed sensitive information.

Future Construction engine (FC): Once the content of email, i.e., header and body part are processed, the future construction engine builds feature sets. The feature sets are the actual source to detect fraudulent emails.

Feature selector (FS): The feature selector uses the concept of tf-idf [4] for word frequency count. Using tf-idf the selected features are further separated into family, business, official, promotion, etc.

Fraudulent email detector: The authors have used WEKA [5] tool to simulate the classification algorithm by feeding email features from feature selector.

Output: The authors calculated the final output using 10-fold cross validation algorithm [6].

2.2 Classification Based on Enhanced Feature Selection [7]

In [7], the authors used the enhanced feature selection strategies along with classification techniques for terrorist email detection. They use various algorithms like Decision Tree [8], Iterative Dichotomiser3 (I.D3) [8], Logistic Regression (LR) [8], Naïve Bayes (NB) [9], and Support Vector Machine (SVM) [8] for detecting email contents comprising frivolous contents. Overall architecture has following modules:

Classification algorithm: The authors use all the above four algorithms for classifying data.

Methodology: The authors evaluated the efficiency of all the classifiers with Feature Selection Strategy (FSS) approach.

Feature selection strategy: FSS is a way to select a subset from original features. The feature selection affects the performance of overall system. High the feature selection, high will be the accuracy and low will be the performance of the overall architecture. The feature selection set (F_s) is a vector of Q and J as: $F_s = \{[Q_1, Q_2 \ldots, Q_N], [J_1, J_2 \ldots, J_N]\}$, Q = Vector of N keywords, and J is vector of Indication. The suspicious feature is defined as J = Js + Jn over Q and Js = suspicious feature and Jn = non-suspicious feature.

System Architecture: The authors have kept feature selection outside main classification. Using WEKA tool they simulated the text message and generated features. The feature functions are kept in repository for further use on different classifications.

Results: The authors has used 10-fold cross validation algorithm and has divided datasets into various combinations of evaluation. The subset of extracted features is applied to all the four classifiers. In each iteration one subset is used to test the efficiency and rest all are used for training. The accuracy function that they used is

$$A = \frac{1}{\mathrm{p}} \sum_{i=1}^{p} ac_i \tag{1}$$

Where ac_i = accuracy of correctly *i*th classified emails in Iteration I, and p is total iterations.

Accuracy after simulation comes out to be as below: (Table 1).

Table 1 Accuracy table of various classifiers

Method	Logistic regression	ID3	Naïve Bayes	SVM linear
Without feature selection	69.64%	78.57%	69.64%	73.21%
CfsSubsetEval, Best First search	83.92%	80.35%	78.57%	80.35%
CfsSubsetEval, Greedy stepwise search	83.92%	83.92%	76.78%	78.57%

3 Ensemble Spam Classifier: Our Approach

After thorough analysis of all the classifications, we come up with an idea of Ensemble Spam classifier. The classifier works in stepwise procedure described as below

Collection of Data: The corpus was downloaded from the mail box using python script.

Initial preprocessing: Unlike other authors, we choose python integrated module called Beautiful Soup to process our raw emails and categorize them as spam and ham.

Creation of Attribute lists and Word list: After preprocessing we calculated the attribute list containing header information like Subject, From, To, Reply To, Recipients, etc. Each mail corresponds to one row in the attribute list (table). The body after preprocessing is stemmed, the words of each mail are stored in numpy array where each row corresponds to one email and overall list is called Word List.

Calculation of frequency of each word in feature vector: The feature vectors contains preprocessed words in utf-8 format, handling strings are cumbersome job, so the attribute list and word list are transformed into integer values using a module Sklearn feature extraction. The sub-module CountVectorizer calculates the local frequencies of each word with respect to given feature vector (each email).

Calculation of global frequencies: As we know, the feature vectors are now transformed into integer values and frequency of each integer (word) is calculated in separate array. There might be many words whose frequencies are same, those words may contribute to false positive. So, in order to remove the discrepancies we calculate the global frequencies based on whole documents (One feature vector will contribute to one document). The tf-idf values are calculated. Tf-idf result of all words will be unique based on the numbers of documents.

Introducing Logistic Regression: As we have binomial class values, i.e., spam and ham. Our classifier has to guess either 0 or 1. If 0, then ham, else spam. The logistic regression is deterministic statistical probabilistic algorithm that works on dichotomous values. So we feed the values to the classifier and give the two class values as 0, 1. The classifier based on the calculated frequencies and training data gives the output.

Results: The logistic regression classifier gives the raw output (inefficient or localized), in order to make the output efficient we map the output of each word in feature vector with a given class, this method is called one versus rest classification (OvR). The result comes out to be as: *One versus rest accuracy: 0.86731,* i.e., the initial efficiency is 86%, which is better than other classifications.

Table 2 Table showing the results on simulating the classifier. Approximately 10000 emails (Ham + Spam) were used to train the classifier. After training, a data set of 2000 emails (Spam + Ham) were used to test the efficiency of the classifier. The results comes out to be as

Classifier	Logistic regression
Static feature vectors	86.731%
Dynamic feature vector(using Genetic Algorithm)	88.931%

Updation using Genetic Algorithm: We recursively choose different attributes and words from the repository to make gradual improvement in our result. Improved result that we got is

One versus rest accuracy: 0.88931, i.e., 89% correctness. This means that our system has overall efficiency of detecting spam as 89% (Table 2).

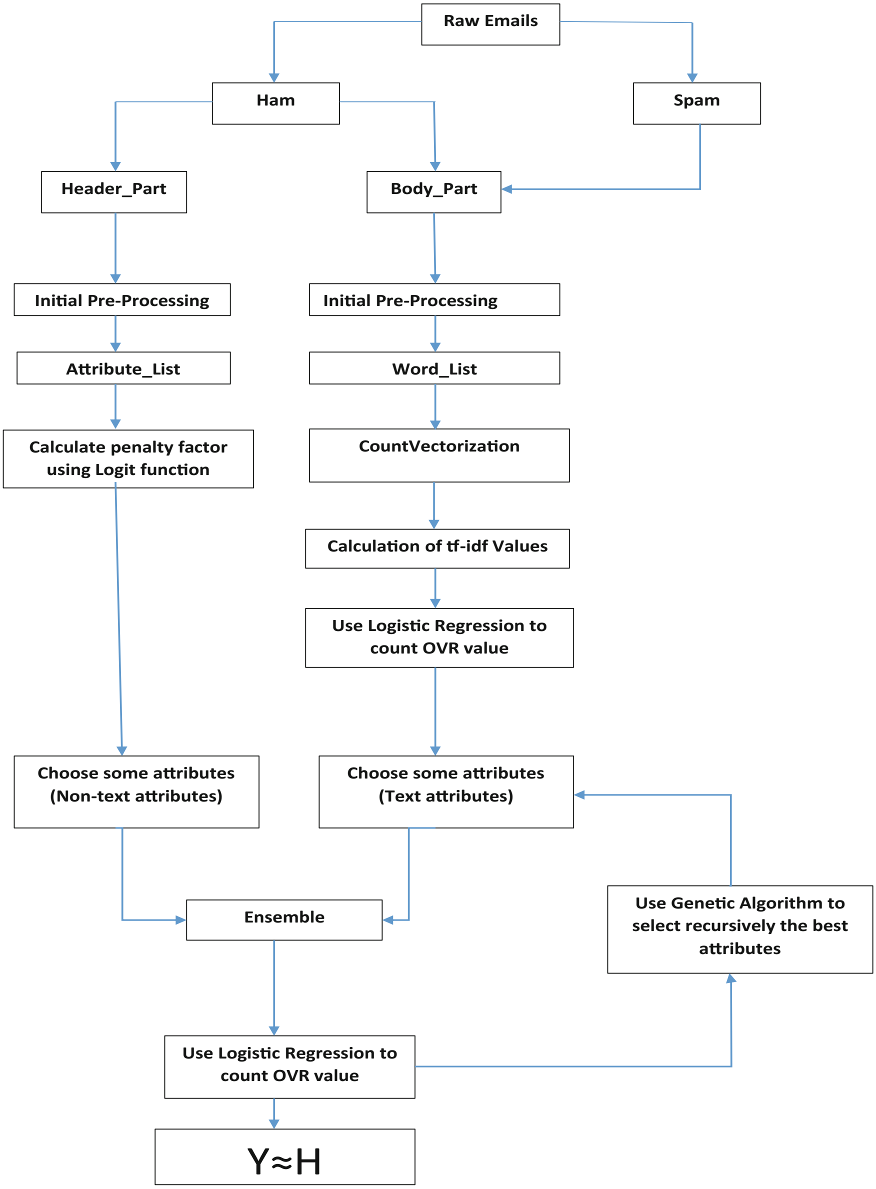

Raw Emails
Ham
Spam
Header_Part
Body_Part
Initial Pre-Processing
Initial Pre-Processing
Attribute_List
Word_List
Calculate penalty factor using Logit function
CountVectorization
Calculation of tf-idf Values
Use Logistic Regression to count OVR value
Choose some attributes (Non-text attributes)
Choose some attributes (Text attributes)
Ensemble
Use Genetic Algorithm to select recursively the best attributes
Use Logistic Regression to count OVR value
Y≈H

Fig. 1 Flow chart of ensemble architecture

The main aim of our proposed work is to make the actual email corpus(Y) equal to that of the predicted one(H). This means that we work to bring the H matrix close to Y matrix. When H matrix comes closer to Y matrix, the classifier starts detecting maximum email correctly as spam and ham. As we increase the size of training corpus, the classifier becomes more and more intelligent. The efficiency of proposed algorithm is highly dependent on the genetic algorithm. We used the deap module to simulate the genetic algorithm and create future generations based on fitness function.

The overall architecture is graphically shown below: (Fig. 1).

4 Conclusion

We studied thoroughly the work based on the classification of emails contents. These classifications are future tools to further strengthen the filtration techniques. The experimental results are based on ideal situations. The work [10] can be studied and improved to great extent. There are other various improvised techniques to classify the email contents like [11, 12]. We studied more classification techniques and came up with the idea of ensemble email classification in which we can modify our work further.

Our work can be further modified using other classifiers. The genetic algorithm can further be trained to improve the accuracy rate.

References

1. N. Sarwat, M. Nasrullah, G. Mathies, and M. Dong Duong Nguyen, "Detection of fraudulent emails by employing advanced feature abundance," Egyptian Informatics Journal (2014) 15, 169–174.
2. S. Nizamani, N. Memon and P. Karampelas, "A text classification model by clustering," International conference on advances in social networks analysis and mining (ASONAM), IEEE; 2011. p. 461–7.
3. S. Nazirova, Survey on Spam Filtering Techniques *Communications and Network* 2011, 3, 153 160.
4. J. Ramos, "Using TF-IDF to determine word relevance in document queries," In: Proceedings of the first instructional conference on machine learning; 2003.
5. M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, IH. Witten, "The WEKA data mining software: an update," ACM SIGKDD Explor Newsl 2009;11(1):10–8.
6. Androutsopoulos, G. Paliouras, V. Karkaletsis, G. Sakkis, C. Spyropoulos, and P. Stamatopoulos, "Machine Learning and Textual Information Access", 4th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD), pages 1–13, 2000.
7. S. Nizamani, Memon, N., Wiil U.K., and Karampelas, P., "Modeling suspicious email detection using enhanced feature selection," Int. J. Modeling and Optimization, vol. 2, pp. 371–377, 2012.

8. S. Nizamani, N. Memon, UK. Wiil and P. Karampelas, “CCM: a text classification model by clustering,” In: 2011 International conference on advances in social networks analysis and mining (ASONAM), IEEE; 2011. p. 461–7.
9. A. McCallum, K. Nigam, “A comparison of event models for Naïve Bayes text classification,” In: AAAI-98 workshop on learning for text categorization, vol. 752; 1998. p. 41–8.
10. W.A. Awad and S.M. ELseuofi, International Journal of Computer Science & Information Technology (IJCSIT), Vol 3, No 1, Feb 2011.
11. Wang, Huiyu, Kai Lei, and Kuai Xu. “Profiling the followers of the most influential and verified users on Sina Weibo”, 2015 IEEE International Conference on Communications (ICC), 2015.
12. S. Puri, D. Gosain, M. Ahuja, I. Kathuria, and N. Jatana, “comparison and analysis of spam detection algorithms,” International Journal of Application or Innovation in Engineering & Management Volume 2, Issue 4, April 2013.

Review Spam Detection Using Opinion Mining

Rohit Narayan, Jitendra Kumar Rout and Sanjay Kumar Jena

Abstract Nowadays with the increasing popularity of Internet, online marketing is going to become more and more popular. This is because, a lot of products and services are easily available online. Hence, reviews about all these products and services are very important for customers as well as organizations. Unfortunately, driven by the will for profit or promotion, fraudsters used to produce fake reviews. These fake reviews written by fraudsters prevent customers and organizations reaching actual conclusions about the products. These fake reviews or review spam must be detected and eliminated so as to prevent deceptive potential customers. In this paper, we have applied supervised learning technique to detect review spam. The proposed work uses different set of features along with sentiment score to build models and their performance were evaluated using different classifiers.

Keywords Review spam · Opinion mining · Sentiment analysis · Machine learning

1 Introduction

Today, because of the popularity of e-commerce sites, spammers have made their target to these sites for review spam apart from other spam like email spam or web spam. Review spam means basically fake review that is written by fraudsters. Mostly e-commerce sites give section for review in order that users can write their opinion about products. There are also many review sites available (like TripAdvisor, Zomato, Yelp etc.) which allow users to write their opinion about the products and

R. Narayan (✉) · J.K. Rout · S.K. Jena
Department of Computer Science & Engineering, National Institute of Technology, Rourkela 769008, Odisha, India
e-mail: rohitnarayan046@gmail.com

J.K. Rout
e-mail: jitu2rout@gmail.com

S.K. Jena
e-mail: skjena@nitrkl.ac.in

P.K. Sa et al. (eds.), *Progress in Intelligent Computing Techniques: Theory, Practice, and Applications*, Advances in Intelligent Systems
and Computing 719, DOI 10.1007/978-981-10-3376-6_30

services. Such type of content provided by web is named as user-generated content. User-generated content contains a lot of valuable and important information about the products and services. Since there is no control on the quality of this content on the web and hence, these promote fraudsters to write fake reviews. These fake reviews prevent customers and organizations reaching actual conclusions about the products. Hence, it highly affects the e-commerce business.

Generally, fake reviews are written for two purposes one for promoting some target objects (positive fake review or positive spam) and another for damage the reputation of other targets (negative fake review or negative spam).

Review spams are generally categorized in three categories [1]: **Untruthful opinions**: If fraudsters write positive fake opinions to promote some targets, it is called as *hyper spam*. If fraudsters write negative fake opinions to damage the reputation of some targets, it is called as *defaming spam*. **Reviews on brand only**: Fraudsters write only about the brand, i.e., the manufacturers of the products rather than the products. **Non-reviews**: Fraudsters write something that is totally unrelated to the products. This may be either advertisements or irrelevant opinion.

Non-reviews can be easily identified manually. Hence in this paper, we have only considered untruthful opinions and reviews on brand only.

The rest of this paper is organized as follows: Sect. 2 presents some surveys related to review spam detection techniques. Section 3 describes the proposed approach for review spam detection. It highlights the datasets, different features, and classifiers those that have been used in our work. Section 4 displays the results obtained using different feature sets. Section 5 gives performance analysis of proposed model and comparison with some existing models and finally Sect. 6 presents conclusion and future directions.

2 Related Work

In the past, a lot of work has been done in the area of spam detection (email spam, web spam, SMS spam). If the sender sends unwanted and unsolicited email either directly or indirectly to user and if there is no relationship of this email to the user, then it is called as email spam. Fette et al. [2] have shown in their work that phishing emails can be easily detected with high accuracy. The proposed work by Li et al. [3] is also based on email spam in which they investigated how to mix multiple email filters supported multivariate analysis so that they can provide a barrier to spam. Web spam refers to the action of the deceptive search engine so that the rank of a specific website becomes more than it deserves [4]. Abernethy et al. [15] provided a graph-based approach for web spam detection. They presented *WITCH* algorithm to detect web spam and also compared this algorithm to many existing algorithms and found that it is better than all those proposed techniques. If someone transmits unwanted and unsolicited messages over communication media (i.e., cell phone), it is called as SMS spam. Karami et al. [5] have used various content features in their work to detect SMS spam.

Detection of opinion spam was first introduced by Jindal and Liu [1] in 2008. They categorized the review spam into three categories (Untruthful opinions, reviews on brands only, and non-reviews) and also built different models for detecting different types of review spam using different supervised learning techniques. A behavioral approach was proposed by Lim et al. [6] to detect review spammers. They tried to find out some behaviors of spammers like they target products and try to maximize their impact. And on the basis of these behaviors they proposed a model to detect review spammers.

The first gold standard positive deceptive opinion spam along with truthful opinion and negative deceptive opinion spam, along with truthful opinion data set for study of review spam, was created by Ott et al. [7, 8]. They have applied some standard features like n-gram and linguistic features for supervised learning to detect fake review and also compare their result with human performance.

Algur et al. [11] in their proposed work used the concept of similarity measure based on conceptual level to detect a given review is spam or non-spam. Feng et al. [12] showed in their work that product reviews contain a natural distribution of opinions and on the basis of this, they built a model to detect review spam. Liu et al. [9] and Mukherjee et al. [10] have used the concept of frequent pattern mining in their work to detect reviewers group. Lim et al. [6] proposed a model that is based on behavior of spammers. They used to assign a rank to spammer on the basis of behavior scoring method and they detect spammers according to that rank.

3 Proposed Approach

We have used three sets of features from different automated approaches along with the sentiment score as a feature. All these features set individual and combination of them are used to train different classifiers and evaluated accuracy. In order to calculate accuracy supervised learning technique has been used in which reviews are divided into two classes, i.e., spam (deceptive review) and non-spam (truthful review). In our work, six classification algorithms, decision tree, naive Bayes, support vector machine (SVM), k-NN, random forest, and logistic regression have been used and taken 80% of the data for training set and 20% of the data for test set with fivefold cross validation.

3.1 Data Set Collection

The dataset that we have used contain 1600 reviews, out of which 800 are truthful reviews (non- spam) containing 400 positive and 400 negative reviews. Similarly 800 deceptive reviews (spam) containing 400 positive and 400 negative reviews. These datasets were created by Ott et al. [7, 8] and public available for research work in the area of review spam detection.

3.2 Features Used

3.2.1 LIWC

The Linguistic Inquiry and Word Count (LIWC) is a text analyzing tool which analyzes 80 different types of features like texts functional aspects, psychological concerns like emotion, perception and personal concerns like money, religion, etc. [14].

3.2.2 POS Tags

Work in linguistics has already proved that the distribution of frequency of parts of speech (POS) tagging of any text is directly dependent on the genre of that text. Hence, according to this approach, feature made for every review is primarily based on the frequency of every POS tag for testing relationship this feature and actual and fake reviews.

3.2.3 N-Gram Feature

In n-gram feature, we select n contiguous words from a text as a feature. If we select one contiguous word at a time as a feature then, it is called as unigram; if we select two contiguous words at a time then, it is bigram and similarly if we select three contiguous words at a time as a features then, it is called as trigram. These features help us to model all the content and its context. In this paper, we have only used unigram as a feature [7].

3.2.4 Sentiment Score

The negative spammers generally used to write more negative words in their review like horrible, disappointed, etc., and hence, show more negative sentiment than a truthful negative review. Similarly, positive spammers used to write more positive words like beautiful, great etc., and show more positive sentiment than an actual positive review. We have calculated sentiment score of each review by the following formula [13]

$$SC(rev) = \sum(-1)^n \frac{S(Wi)}{Distance(Wi, fet)}$$

Here,
rev is review text
fet is aspect/feature in a sentence of review
S(Wi): sentiment polarity of Wi word (+1 or -1)
n: total number of negation words in a feature, default = 0 and
Distance (Wi, fet): distance between word Wi and feature fet.

4 Results and Discussions

Our experimental results are based on dataset described in Sect. 3. All the features discussed in previous section have been used individually as well as in some combinations to train different classifiers. In all the cases, logistic regression classifier gives better result. If only LIWC features used, maximum accuracy achieved by logistic regression classifier is 61.87%. LIWC features along with sentiment score gives maximum accuracy of 68.75%. Only POS features give maximum accuracy of 65.62%. LIWC features along with POS features gives maximum accuracy of 70.62%. If we combine both LIWC and POS features along with sentiment score, maximum accuracy obtained is 76.25%. Only unigram as feature gives maximum accuracy of 75.62% however, if combined with sentiment score, maximum accuracy achieved is up to 81.25 %. Overall maximum accuracy obtained is 86.25% by combining LIWC and unigram along with sentiment score as features for Logistic Regression classifier. The results have been shown in Table 1.

5 Performance Analysis

Algur et al. [11] used the concept of similarity measure based on conceptual level to detect a given review is spam or not. They have shown that their proposed technique gives the accuracy of 57.29% for pros reviews and 30.00% for cons reviews or average 43.64% accuracy. Ott et al. [7] noticed that human can determine review spam manually with maximum accuracy of 61.9%. Feng et al. [12] observed in their work that product reviews contain a natural distribution of opinions and on the basis of this they build a model to detect review spam with accuracy 72.5%. Lim et al. [6] proposed a model that is based on behavior of spammers. They used to assign a rank to spammers on the basis of behavior scoring method and according to that rank

Table 1 Different features analysis results

Approaches	Maximum accuracy (%)	Precision	Recall	F-score
LIWC	61.87	57.50	63.01	60.13
LIWC, sentiment score	68.75	60.00	72.72	65.75
POS	65.62	67.50	65.06	66.25
LIWC, POS	70.62	66.25	72.60	69.28
LIWC, POS, sentiment score	76.25	77.50	75.60	76.54
Unigram	75.62	77.50	74.69	76.07
Unigram, sentiment score	81.25	90.00	76.59	82.75
Unigram, LIWC, sentiment score	86.25	90.00	83.72	86.72

Table 2 Comparative performance analysis

Method	Accuracy (%)
Conceptual level similarity measure based review spam detection [11]	43.64
Review spam detection by human performance [7]	61.90
Distributional footprints of deceptive product reviews [12]	72.50
Detecting product review spammers using rating [6]	78.00
Proposed approach	86.25

they detect spammers with accuracy of 78%. Our proposed approach gives maximum accuracy of 86.25%. Based on accuracy a comparative result is shown in Table 2.

6 Conclusion

In this work, we have used three sets of features, i.e., LIWC, POS, and n-gram from different automated approaches along with the sentiment score. These feature sets have been used individually as well as in some combinations to train different classifiers. Six classification algorithm were employed such as: decision tree, naive Bayes, SVM, k-NN, random forest, and logistic regression. Our experimental results reveals that logistic regression outperforms other classifiers. In the case of individual feature set, unigram gives maximum accuracy of 75.62% with F-score 76.07. However, for combinations unigram and LIWC along with sentiment score gives accuracy of 86.25% with F-score 86.72 and that is maximum. At last, we have compared our proposed technique with some existing review spam detection techniques on the basis of their accuracy which shows our technique gives better result than others. In this work, we have used supervised learning method and in future the same work can be extended for semi-supervised learning as well as unsupervised learning method to overcome the unavailability of labeled data sets.

References

1. Jindal, Nitin, and Bing Liu.: Opinion spam and analysis. In: Proceedings of the 2008 International Conference on Web Search and Data Mining. ACM, 2008.
2. Fette, Ian, Norman Sadeh, and Anthony Tomasic.: Learning to detect phishing emails. In: Proceedings of the 16th international conference on World Wide Web. ACM, 2007.
3. Li, Wenbin, Ning Zhong, and Chunnian Liu.: Combining multiple email filters based on multivariate statistical analysis. In: Foundations of Intelligent Systems. Springer Berlin Heidelberg, 2006. 729–738.
4. Spirin, Nikita, and Jiawei Han.: Survey on web spam detection: principles and algorithms. In: ACM SIGKDD Explorations Newsletter 13.2 (2012): 50–64.

5. Karami, Amir, and Lina Zhou.: Improving static SMS spam detection by using new content-based features. (2014).
6. Lim, Ee-Peng, et al.: Detecting product review spammers using rating behaviors. In: Proceedings of the 19th ACM international conference on Information and knowledge management. ACM, 2010.
7. Ott, Myle, et al.: Finding deceptive opinion spam by any stretch of the imagination. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Association for Computational Linguistics, 2011.
8. Ott, Myle, Claire Cardie, and Jeffrey T. Hancock.: Negative Deceptive Opinion Spam. In: HLT-NAAC L. 2013.
9. Mukherjee, Arjun, et al.: Detecting group review spam. In: Proceedings of the 20th international conference companion on World wide web. ACM, 2011.
10. Mukherjee, Arjun, Bing Liu, and Natalie Glance.: Spotting fake reviewer groups in consumer reviews. In: Proceedings of the 21st international conference on World Wide Web. ACM, 2012.
11. Algur, Siddu P., et al.: Conceptual level similarity measure based review spam detection. In: Signal and Image Processing (ICSIP), 2010 International Conference on. IEEE, 2010.
12. Feng, Song, et al.: Distributional Footprints of Deceptive Product Reviews. In: ICWSM 12 (2012): 98-105.
13. Peng, Qingxi, and Ming Zhong.: Detecting spam review through sentiment analysis. In: Journal of Software 9.8 (2014): 2065-2072.
14. Harris, C.: Detecting deceptive opinion spam using human computation. In: Workshops at AAAI on Artificial Intelligence. 2012.
15. Abernethy, Jacob, Olivier Chapelle, and Carlos Castillo.: Graph regularization methods for web spam detection. Machine Learning 81.2 (2010): 207-225.

Review Spam Detection Using Semi-supervised Technique

Rohit Narayan, Jitendra Kumar Rout and Sanjay Kumar Jena

Abstract Today because of the popularity of e-commerce sites, spammers have made their target to these sites for review spam apart from other spams like email spam or web spam. These fake reviews written by fraudsters prevent customers and organizations reaching actual conclusions about the products. Hence, these must be detected and eliminated so as to prevent deceptive potential customers. In this paper, we have used semi-supervised learning technique to detect review spam. The proposed work is based on PU-learning algorithm which learns from a very few positive example and unlabeled data set. Maximum accuracy we have achieved is of 78.12% with F-score 76.67 using only 80 positive example as a training set.

Keywords Review spam · Semi-supervised learning technique · Opinion spam

1 Introduction

Nowadays with the increasing popularity of Internet, online marketing is going to become more and more popular. This is because, a lot of products and services are easily available online. Hence, reviews about these all products and services are very important for customers as well as organizations. Because of this, spammers have made their target to these sites for review spam apart from other spam like email spam or web spam. Mostly e-commerce sites give section for review in order that users can write their opinion about products. There are also many review sites available (like TripAdvisor, Yelp, etc.) which allow users to write their opinion about

R. Narayan (✉) · J.K. Rout · S.K. Jena
Department of Computer Science & Engineering, National Institute of Technology,
Rourkela 769008, Odisha, India
e-mail: rohitnarayan046@gmail.com

J.K. Rout
e-mail: jitu2rout@gmail.com

S.K. Jena
e-mail: skjena@nitrkl.ac.in

P.K. Sa et al. (eds.), *Progress in Intelligent Computing Techniques: Theory, Practice, and Applications*, Advances in Intelligent Systems
and Computing 719, DOI 10.1007/978-981-10-3376-6_31

the products and services. Such type of content provided by web is named as user-generated content. User-generated content contains a lot of valuable and important information about the products and services. Since there is no control on the quality of this content on the web and hence, these promote fraudsters to write fake and wrong information about the products. These fake and wrong information written by fraudsters is called as review spam. Fake reviews prevent customers and organizations reaching actual conclusions about the products. Hence, it highly affects the e-commerce business. That is why, over the last few years, these review sites have been removing fake reviews about from their websites using their own spam detection technique.

Generally, fake reviews are written for two purposes one for promoting some target objects (positive fake review or positive spam) and another for damage the reputation of other targets (negative fake review or negative spam). Machine-learning techniques have been more popular for spam detection. They uses supervised (required all data set labeled), semi-supervised (require very few data set labeled) and unsupervised (works for unlabeled data set) learning technique.

The rest of this paper is organized as follows: Sect. 2 presents some surveys related to review spam detection techniques. Section 3 describes the proposed approach for review spam detection. It highlights the data sets, and semi-supervised technique that has been used in our work. Section 4 displays the results obtained using proposed technique and finally Sect. 5 presents conclusion and future directions.

2 Related Work

In the past, a lot of work has been done in the area of spam detection (email spam, web spam, SMS spam). If the sender sends unwanted and unsolicited email either directly or indirectly to user and there is no relationship of this email to the user is called as email spam. Fette et al. [1] have shown in their work that phishing emails can be easily detected with high accuracy. The proposed work by Li et al. [2] is also based on email spam in which they investigated how to mix multiple email filters supported multivariate analysis so that they can provide a barrier to spam. Web spam refers to the action of the deceptive search engine so that the rank of a specific website becomes more than it deserves [3]. Abernethy et al. [4] provided a graph-based approach for web spam detection. They presented *WITCH* algorithm to detect web spam and also compared this algorithm to many existing algorithms and found that it is better than all those proposed techniques. If someone transmits unwanted and unsolicited messages over communication media (i.e., cell phone) is called as SMS spam. Karami et al. [5] have used various content features in their work to detect SMS spam.

Detection of opinion spam was first introduced by Jindal and Liu [6] in 2008. They categorized the review spam into three categories (Untruthful opinions, reviews on brands only, and non-reviews) and also built different models for detecting different types of review spam using different supervised learning techniques. The first gold standard positive deceptive opinion spam along with truthful opinion and negative deceptive opinion spam along with truthful opinion data set for study of review spam was created by Ott et al. [7, 8]. They have applied some standard features like n-gram and linguistic features for supervised learning to detect fake review and also compare their result with human performance.

A lot of work has been done in supervised learning technique. But the drawback is we need to label all the data set. To overcome such problem, Hernndez et al. [9] applied the concept of semi-supervised learning technique to detect review spam detection. Authors used the data set created by Ott et al. [7] containing 800 positive reviews out of which 400 are deceptive and 400 are truthful. They took 160 data set as a test set which is labeled and for training took 520 unlabeled data set and combination of 20, 40, 60, 80, 100, 120 as a positive instances. After that they applied PU-learning algorithm on these positive and unlabeled instances to calculate accuracy. They used one-class, naive bayes, and SVM classifier in their work.

Liu et al. [10] in their proposed work also used the concept of semi-supervised learning technique to detect spam. They divided data set into two set of classes. A particular data set comes into class P, a large number of data set come in class M. Such technique is called as partially supervised classification. They used EM algorithm to identify class P from class M. EM algorithm generates a sequence of solutions. For each solution they used naive bayes classifier to calculate accuracy.

3 Proposed Approach

Semi-supervised learning technique is a machine-learning technique that uses a large amount of unlabeled data and a very few labeled data set for training. Semi-supervised learning lies between supervised learning (completely labeled data) and unsupervised learning (completely unlabeled data). Many researchers found that if a large amount of unlabeled data, when used with a few labeled data set, can produce good accuracy in term of learning problem.

3.1 Data set Description

The data set that we have used contain 800 positive opinion reviews in combination of truthful and deceptive. These data sets were created by Ott et al. [7] and public available for research work in the area of review spam detection.

3.2 Proposed Model

In proposed method, we have taken different sub-corpa from data sets. For building test set, first we randomly selected 160 opinions, out of which 80 are deceptive and 80 are truthful. The rest 640 opinions have been used for three different size of training sets. They consist 40, 80, and 120 positive instances (deceptive opinion) respectively. In all the cases, we have used 520 unlabeled instances. Now, PU-learning algorithm [9] has been used for review spam detection.

Algorithm 1: PU-Learning for Spam Detection

```
$i \leftarrow 1$;
$|W_0| \leftarrow |U_1|$;
$|W_1| \leftarrow |U_1|$;
while $|W_i| \leq |W_{i-1}|$ do
    $C_i \leftarrow Generate\_Classifier(P, U_i)$;
    $U_i^L \leftarrow C_i(U_i)$;
    $W_i \leftarrow Extract\_Positives(U_i^L)$;
    $U_{i+1} \leftarrow U_i$ - $W_i$;
    $i \leftarrow i + 1$;
Return Classifier $C_i$;
```

here,

P: Set of positive instances.
U_i: Unlabeled set at iteration i.
U_1: Original unlabeled data set.
C_i: Classifier at iteration i.
W_i: Unlabeled instance classified as positive by classifier C_i.

Proposed technique is an iterative process in which first all unlabeled instance are considered as part of negative class. After that we have used positive instances to train different classifiers. Here, six classifiers have been used. These are decision Tree, naive Bayes, support vector machine, k-NN, random forest, and logistic regression. After that we classify unlabeled data set using these classifiers. All positive instances are eliminated from instances of unlabeled data and rest are treated as negative instance for next iteration. This process is repeated until a stop criteria is achieved.

4 Results and Discussions

Our experimental results using semi-supervised learning technique are based on data set described in Sect. 5. For different training sets result has been shown in Tables 1, 2 and 3.

Table 1 Results of different classifiers when using 40 deceptive opinions as training and 520 unlabeled opinions

Training sets	Classifiers	Accuracy (%)	P	R	F
Positive instance = 40 Unlabeled set = 520	Decision tree	56.25	56.25	56.25	56.25
	Naive Bayes	40.62	12.50	28.57	17.39
	SVM	54.68	87.50	52.83	65.88
	k-NN	**64.06**	75.00	61.53	67.60
	Random forest	54.68	71.87	53.48	61.33
	Logistic regression	60.93	78.12	58.13	66.67

Table 2 Results of different classifiers when using 80 deceptive opinions as training and 520 unlabeled opinions

Training Sets	Classifiers	Accuracy (%)	P	R	F
Positive instance = 80 Unlabeled set = 520	Decision tree	70.31	65.62	72.41	68.85
	Naive Bayes	56.25	25.00	66.67	36.36
	SVM	71.87	87.50	66.67	75.67
	k-NN	**78.12**	71.87	82.14	76.67
	Random forest	65.62	56.25	69.23	62.06
	Logistic regression	76.56	84.37	72.97	78.26

Table 3 Results of different classifiers when using 120 deceptive opinions as training and 520 unlabeled opinions

Training sets	Classifiers	Accuracy (%)	P	R	F
Positive instance = 120 Unlabeled set = 520	Decision tree	45.31	50.00	45.71	47.76
	Naive Bayes	54.68	34.37	57.89	43.13
	SVM	60.93	90.62	56.86	69.87
	k-NN	60.93	71.87	58.97	64.78
	Random forest	46.87	56.25	47.36	51.42
	Logistic regression	**73.43**	68.75	75.86	72.13

Hence, maximum accuracy we have achieved is of 78.12 % with F-score 76.67 when used 80 examples of deceptive opinions from data sets as training set with 520 unlabeled data set and 160 labeled data as a test set using k-NN classifier.

5 Conclusion

In this work, we applied PU-learning algorithm along with six different classifiers (decision tree, naive bayes, SVM, k-NN, random forest, logistic regression) to detect review spam from our data set. We have taken different sub-corpa from data sets. For building test set, first we randomly selected 160 opinions, out of which 80 are deceptive and 80 are truthful. The rest 640 opinions have been used for three different sizes of training sets. They consist 40, 80, and 120 positive instances (deceptive opinion) respectively. In all the cases, we have used 520 unlabeled instances. Now, PU-learning algorithm has been used for review spam detection. Maximum accuracy we have achieved is of 78.12 % with F-score 76.67 when used 80 examples of deceptive opinions from data sets as training set with 520 unlabeled data set using k-NN classifier. In future, the same work can be extended for unsupervised learning technique to overcome the unavailability of labeled data sets.

References

1. Fette, Ian, Norman Sadeh, and Anthony Tomasic.: Learning to detect phishing emails. In: Proceedings of the 16th international conference on World Wide Web. ACM, 2007.
2. Li, Wenbin, Ning Zhong, and Chunnian Liu.: Combining multiple email filters based on multivariate statistical analysis. In: Foundations of Intelligent Systems. Springer Berlin Heidelberg, 2006. 729–738.
3. Spirin, Nikita, and Jiawei Han.: Survey on web spam detection: principles and algorithms. In: ACM SIGKDD Explorations Newsletter 13.2 (2012): 50–64.
4. Abernethy, Jacob, Olivier Chapelle, and Carlos Castillo.: Graph regularization methods for web spam detection. Machine Learning 81.2 (2010): 207–225.
5. Karami, Amir, and Lina Zhou.: Improving static SMS spam detection by using new content-based features. (2014).
6. Jindal, Nitin, and Bing Liu.: Opinion spam and analysis. In: Proceedings of the 2008 International Conference on Web Search and Data Mining. ACM, 2008.
7. Ott, Myle, et al.: Finding deceptive opinion spam by any stretch of the imagination. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Association for Computational Linguistics, 2011.
8. Ott, Myle, Claire Cardie, and Jeffrey T. Hancock.: Negative Deceptive Opinion Spam. In: HLT-NAAC L. 2013.
9. Hernndez, D., et al.: Using PU-learning to detect deceptive opinion spam.: Proc. of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis., 2013.
10. Liu, Bing, et al.: Partially supervised classification of text documents. ICML. Vol. 2. 2002.

Part V
Emerging Techniques in Computing

Dimensionality Reduction Using Decision-Based Framework for Classification: Sky and Ground

Ramesh Ashok Tabib, Ujwala Patil, T. Naganandita, Vinita Gathani and Uma Mudenagudi

Abstract In this paper, we address the problem of dimensionality reduction for classification. Classification of data is challenging if its dimension size is high. We propose a decision-based framework for dimensionality reduction using confidence factor as an evaluation measure for generating a relevant feature subset for a specific target. Confidence factor is generated for all features competent for classification using evidence parameters. Evidence parameters are computed based on intersection of classes in the distribution of feature vector and distance between peaks of distribution of feature vector and are combined using Dempster Shafer combination rule. We demonstrate the results of the proposed framework for sky and ground classification using various datasets. The classification in low dimension space is performed retaining the classification accuracy and optimizing computational time.

Keywords Dimensionality reduction (DR) · Competent feature vector set s · Optimal feature vector set r · Evidence parameter · Dempster Shafer combination rule (DSCR) · Confidence factor

R.A. Tabib (✉) · U. Patil · T. Naganandita · V. Gathani · U. Mudenagudi
Department of Electronics and Communication Engineering, BVBCET,
Hubli 580031, India
e-mail: ramesh_t@bvb.edu

U. Patil
e-mail: ujwalapatil@bvb.edu

T. Naganandita
e-mail: t.naganandita@gmail.com

V. Gathani
e-mail: gathanivinita@gmail.com

U. Mudenagudi
e-mail: uma@bvb.edu

P.K. Sa et al. (eds.), *Progress in Intelligent Computing Techniques: Theory, Practice, and Applications*, Advances in Intelligent Systems
and Computing 719, DOI 10.1007/978-981-10-3376-6_32

1 Introduction

Real-world data is observed to have a high-dimensional space and requires sophisticated methods for classification and analysis. The computational cost of the classification of data increases quadratically with the increase in dimension size [1], which is referred as the *curse of dimensionality*. One of the approaches to address curse of dimensionality is to reduce the dimensions of the data through process of dimensionality reduction (DR).

Dimensionality reduction is the process of meaningful representation of high-dimensional data in lower dimensional space. The optimal dimensional data ideally refers to the intrinsic dimensionality of the data, which is the minimum number of dimensions required to represent the observed properties of high-dimensional data [15]. DR is important in many fields as it facilitates compression, visualization, and classification to eliminate undesired properties of the high dimensional data.

Several reduction techniques are proposed in the literature for DR [5, 8, 9, 11, 14]. They are mainly classified into linear [9, 11] and nonlinear [4, 7] methods. Earlier linear techniques were used to perform DR. As the complex nonlinear data is not adequately handled by the linear techniques, various nonlinear techniques are proposed since last decade. These techniques offer major advantages in real-world data as it is highly non linear. Even though the nonlinear techniques are successful in handling the artificial datasets, they are not capable of producing satisfactory results on the natural datasets [15].

Dimensionality reduction plays an important role in the field of machine learning and classification. Classification of high-dimensional data is challenging. Competent feature vectors are the features that give the best performance under some classification system.

Feature selection/elimination is mainly used in the areas of clustering and classification [10]. In data processing, feature selection/elimination is one of the most frequently used techniques as it reduces the number of features to be collected for classification and improves the speed of algorithms. In addition to this, it may also result in a better classification accuracy.

In this paper, we propose a nonlinear dimensionality reduction technique using decision-based feature elimination.

The main contributions of the paper are:

1. We propose a decision-based framework for dimensionality reduction using confidence factor as an evaluation measure for generating a relevant feature vector subset for a specific target.
2. We propose to generate evidence parameters based on intersection area between the classes in distribution of feature vector and distance between peaks of distribution of feature vector.
3. We generate confidence factor by combining evidence parameters using DSCR.

We present the proposed framework in Sect. 2, discussion of results in Sect. 3 and conclusions in Sect. 4.

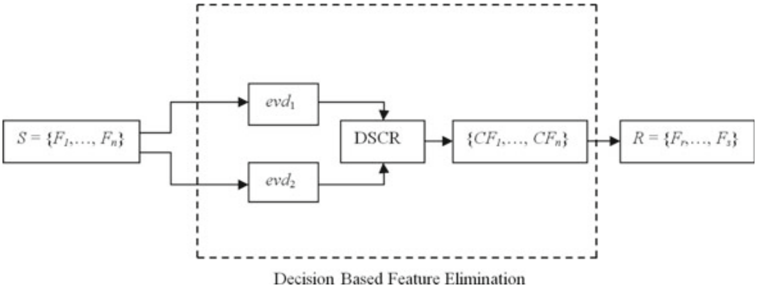


Fig. 1 Framework for Dimensionality Reduction

2 Decision-Based Framework for Dimensionality Reduction

In this section, we propose a decision-based feature elimination technique for dimensionality reduction. If U is the universal set of feature vectors, we choose set of feature vectors S $\{F_1, \ldots, F_n\}$ which is a subset of universal set U and considered to be a competent feature vector set sufficient for a specific classification problem. Some of the competent feature vectors are sufficient but not optimal for addressing the classification problem. Proposed framework identifies the optimal feature vectors for specific classification problem as explained in Sect. 2.1. The reduction in the number of competent feature vectors for classification addresses the problem of dimensionality reduction (Fig. 1).

2.1 *Decision-Based Feature Elimination*

Let F_p be the pth feature vector of the competent feature vector set S used to classify the two set classes $\{C_1, C_2\}$. Let $\{F_r, \ldots, F_s\}$ be the optimal feature vector set R of optimal competent feature vectors which is a subset of S. The element F_p is included in subset R if its confidence factor is higher than or equal to a set threshold $\mathcal{T}$ otherwise it is eliminated. Confidence factor is generated from evidence parameters. Evidence parameters are computed based on intersection of classes in the distribution of feature vector and distance between peaks of distribution in feature vector and are combined using Dempster–Shafer combination rule (DSCR).

2.2 *Generation of Evidence Parameters*

We consider two evidence parameters evd_1 and evd_2. Evidence parameter evd_1 is generated based on intersection area of classes in the distribution of feature vector.

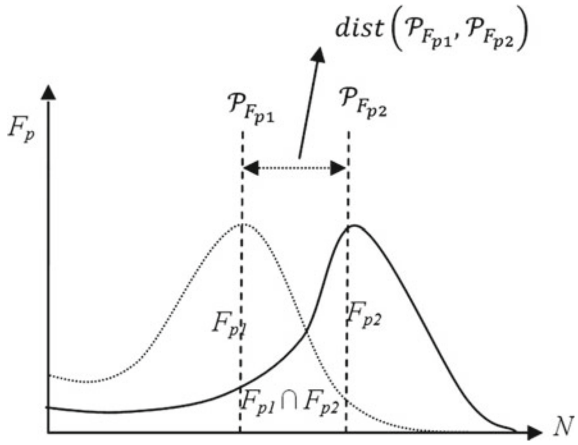


Fig. 2 Distribution of feature vector F_p

Evidence parameter evd_2 is computed based on the distance between peaks of distribution of feature vector. Evidence parameters evd_1 and evd_2 are calculated as shown in Eqs. 1 and 2, respectively.

Let F_{p1} and F_{p2} be the distribution of feature vectors for class C_1 and C_2 respectively. The distribution of feature vectors F_{p1} and F_{p2} is shown in Fig. 2. Intersection area of F_{p1} and F_{p2} is used to compute the evidence parameter evd_1 toward the confidence factor CF_p for the feature vector F_p and is given by

$$evd_1 = 1 - norm(F_{p1} \cap F_{p2}) \tag{1}$$

where the intersection area is normalized with respect to the total area of classes in the distribution of feature vector.

The distance between peaks $\mathcal{P}_{F_{p1}}$ and $\mathcal{P}_{F_{p2}}$ of feature vectors F_{p1} and F_{p2} is used to compute the evidence parameter evd_2 toward the confidence factor CF_p for the feature F_p and is given by

$$evd_2 = norm(dist(\mathcal{P}_{F_{p1}}, \mathcal{P}_{F_{p2}}) \tag{2}$$

where the distance is normalized with respect to the total length of the distribution of feature vector.

2.3 Dempster Shafer Combination Rule (DSCR)

We combine evidence parameters using DSCR to generate confidence factor. The evidence parameters explained in Sect. 2.2 are considered as the mass of belief function for DSCR [6]. We propose to use confidence factor as thresholding parameter for generation of optimal feature vector set R. Let hyp_1 and hyp_2 be the hypothesis supporting the confidence for belief and disbelief towards the feature vector to be a part of optimal feature vector set R. We denote the complete system in a set as, $2^{\mathcal{A}} = \{\emptyset, \{hyp_1\}, \{hyp_2\}, \mathcal{A}\}$, where $\mathcal{A} = \{hyp_1, hyp_2\}$ denotes the state of ambiguity and $\emptyset$ is considered to be the conflict between the set hypothesis. Dempster–Shafer proposed a rule of combination [3, 6, 13] using which we combine the evidence as masses $m(evd_1)$ and $m(evd_2)$. Let $\mathcal{H}$ denote the combined hypothesis, then according to the DSCR [6], the numerators of Eqs. 3 and 4 represent the accumulated/collected evidences from evd_1 and evd_2, which are in the favor of combined hypothesis $\mathcal{H}$. Summation term in denominator of Eq. 3 is considered to be the mass of conflict in the combined hypothesis. The denominator in Eq. 4 specifies about the collective mass of belief, disbelief, and ambiguity towards the hypothesis, which acts as the normalizing factor toward the set hypothesis.

The computed mass of combined hypothesis $\mathcal{H}$ is given by,

$$m(\mathcal{H}) = \frac{\sum_{evd_1 \cap evd_2 = \mathcal{H} \neq \phi} m(evd_1).m(evd_2)}{1 - \sum_{evd_1 \cap evd_2 = \phi} m(evd_1).m(evd_2)} \tag{3}$$

which can be written as:

$$m(\mathcal{H}) = \frac{\sum_{evd_1 \cap evd_2 = \mathcal{H} \neq \phi} m(evd_1).m(evd_2)}{\sum_{evd_1 \cap evd_2 \neq \phi} m(evd_1).m(evd_2)} \tag{4}$$

The subset considered for contribution toward combined belief of set hypothesis is:

$\{hyp_1\} = \{m(\mathcal{H}_{11}), m(\mathcal{H}_{13}), m(\mathcal{H}_{31})\} = \{m(K_1), m(K_2), m(K_3)\}$.

Similarly, the subset considered for contribution toward combined disbelief of set hypothesis is:

$\{hyp_2\} = \{m(\mathcal{H}_{22}), m(\mathcal{H}_{23}), m(\mathcal{H}_{32})\} = \{m(K_4), m(K_5), m(K_6)\}$.

The subset considered for contribution towards both belief and disbelief is null set,
$\{\emptyset\} = \{m(\mathcal{H}_{12}), m(\mathcal{H}_{21})\}$.

The set considered to be ambiguous is, $\{\mathcal{A}\} = \{m(\mathcal{H}_{33})\} = \{m(K_7)\}$.

The masses of hypothesis $\{hyp_1\}$ and $\{hyp_2\}$ as per Eq. 4 are given by:

$$m(\{hyp_1\}) = \frac{\sum_{i=1}^{3} m(K_i)}{\sum_{i=1}^{7} m(K_i)} \tag{5}$$

$$m(\{hyp_2\}) = \frac{\sum_{i=4}^{6} m(K_i)}{\sum_{i=1}^{7} m(K_i)} \tag{6}$$

The masses $m(\{hyp_1\})$ and $m(\{hyp_2\})$ are the values of confidence factors in favor and against the consideration of feature vector F_p as optimal feature vector, respectively (Table 1).

Table 1 The hypothesis combination table shows the plot of combined hypothesis $\mathcal{H}$ using evd_1 and evd_2 [6, 12]

$\cap$	$m(evd_1^{belief})$	$m(evd_1^{disbelief})$	$m(evd_1^{ambiguity})$
$m(evd_2^{belief})$	$hyp_1 \leftarrow m(\mathcal{H}_{11})$	$\emptyset$	$hyp_1 \leftarrow m(\mathcal{H}_{13})$
$m(evd_2^{disbelief})$	$\emptyset$	$hyp_2 \leftarrow m(\mathcal{H}_{22})$	$hyp_2 \leftarrow m(\mathcal{H}_{23})$
$m(evd_2^{ambiguity})$	$hyp_1 \leftarrow m(\mathcal{H}_{31})$	$hyp_2 \leftarrow m(\mathcal{H}_{32})$	$\mathcal{A} \leftarrow m(\mathcal{H}_{33})$

We assign the mass of hypothesis hyp_1, $m(\{hyp_1\})$ as the confidence factor, CF_p for feature vector F_p.

2.4 Generation of Optimal Set R

The competent feature vector F_p belongs to the set of optimal feature vectors R if the confidence factor CF_p lies above a set threshold $\mathcal{T}$ as shown in Eq. 7.

$$R = \begin{cases} Fp \in R & \text{if } CFp > \mathcal{T}, \\ Fp \notin R & \text{otherwise.} \end{cases} \tag{7}$$

This optimal feature vector set R is used for testing the input data.

3 Results and Discussions

In this section, we discuss the results of the proposed framework using four datasets of various sky and ground conditions. We select texture features as competent feature vectors for classification of sky and ground. The texture features considered are intensity, mean intensity, standard deviation of intensities, third moment, smoothness, energy and entropy [16] for all three channels in the image. In our experiments, we observe that the optimal feature vector set R varies in accordance with the dataset as shown in Table 4. We demonstrate that the computational time for classification is optimized retaining the accuracy of classification.

3.1 Decision-Based Feature Elimination

The competent feature vector set S considered for sky and ground classification consists of 21 features, i.e., the seven feature vectors in each channel as mentioned above. For demonstrating the results using competent feature vector set S and the optimal

Table 2 Confidence factor for feature vectors F_1 to F_{11}

	Feature vectors										
	F_1	F_2	F_3	F_4	F_5	F_6	F_7	F_8	F_9	F_{10}	F_{11}
Dataset I	70	64	16	0	0	0	0	7	17	40	8
Dataset II	99	29	84	98	0	0	0	0	0	0	0
Dataset III	99	67	99	99	0	0	0	99	39	57	56
Dataset IV	0	0	0	0	0	0	0	76	97	33	0

Table 3 Confidence factor for feature vectors F_{12} to F_{21}

	Feature vectors									
	F_{12}	F_{13}	F_{14}	F_{15}	F_{16}	F_{17}	F_{18}	F_{19}	F_{20}	F_{21}
Dataset I	0	0	0	90	38	41	0	0	0	0
Dataset II	0	0	0	0	0	0	0	0	0	0
Dataset III	0	0	0	65	31	45	0	0	0	0
Dataset IV	0	0	0	97	98	24	0	0	0	0

Table 4 Optimal feature vector set

Dataset	$\mathcal{T}$	Optimal feature vector set R	No. of feature vectors in R
I	90	$\{F_{15}\}$	1
II	80	$\{F_1, F_3, F_4\}$	3
III	90	$\{F_1, F_3, F_4, F_8\}$	4
IV	75	$\{F_8, F_9, F_{15}, F_{16}\}$	4

feature vector set R, we take the samples of sky and ground in the twilight, morning, afternoon and evening, and divide them into four datasets namely Dataset I, Dataset II, Dataset III, and Dataset IV respectively. The confidence factor for each competent feature vector in every dataset is calculated using the above framework and the values are shown in Tables 2 and 3.

We set the value of threshold $\mathcal{T}$ heuristically for different datasets. The feature vectors having confidence factor greater than or equal to threshold $\mathcal{T}$ are included in the optimal feature vector set R. The optimal feature vector set R for each dataset is as shown in Table 4.

The number of feature vectors in the optimal set R varies for a set threshold $\mathcal{T}$ as shown in Table 4. In dataset I, the value of $\mathcal{T}$ is set to 90 and only one feature vector is selected to the optimal set R whereas for the same value of $\mathcal{T}$, four feature vectors are selected for dataset III. Therefore, the value of $\mathcal{T}$ has to be set experimentally for optimum classification.

It is evident from the Table 4 that there is a drastic reduction in the number of competent feature vectors required for classification. This demonstrates the dimensionality reduction achieved from the proposed framework.

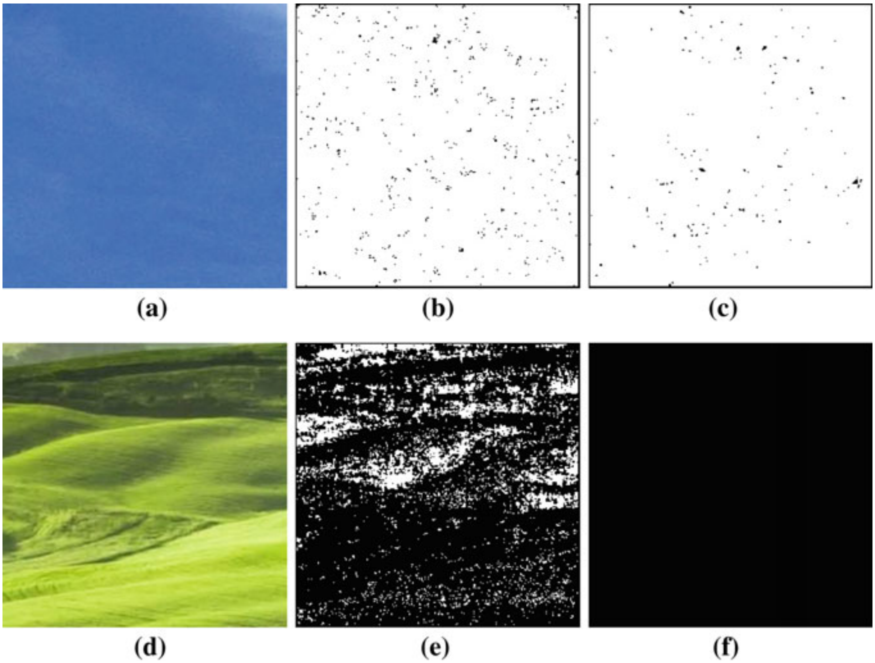


Fig. 3 Classification results for samples of sky and ground from Dataset II **a** Input image. **b** Classified using set *S*. **c** Classified using set *R*. **d** Input image. **e** Classified using set *S*. **f** Classified using set *R*

3.2 Qualitative Quality Analysis of Classification

We use a variant of Bayesian classifier for classification [2]. The number of samples of each dataset used for training is 2.6 million and for testing is 1.3 million. The samples of each dataset are tested using the competent feature vector set *S* and the optimal feature vector set *R*. We set an intensity value of 255 if the sample is classified as sky otherwise it is set to 0. The results of testing are shown in Fig. 3.

3.3 Quantitative Quality Analysis of Classification

The computational time and classification accuracy for the feature vector sets *S* and *R* are shown in Tables 5 and 6.

For dataset I, the classification accuracy decreases by 6% but there is a significant reduction (94%) in computational time as the number of feature vectors in the optimal set is only one. For dataset II, the classification accuracy increases by 5% as the ambiguity in classification using competent feature vectors with less confidence

Table 5 Computational time for classification using competent feature vector set *S* and optimal feature vector set *R*

Dataset	Computational time with set *S* (s)	Computational time with set *R* (s)	Percentage decrease in time
I	1090	66	94
II	1100	220	80
III	1095	275	75
IV	1092	230	79

Table 6 Classification accuracy for classification using competent feature vector set *S* and optimal feature vector set *R*

Dataset	Classification accuracy with set *S* (%)	Classification accuracy with set *R* (%)	Percentage change in accuracy
I	96	90	Decrease by 6%
II	93	98	Increase by 5%
III	99	98	Decrease by 1%
IV	94	96	Increase by 2%

is higher than the feature vectors of the optimal set. In dataset III and IV, the classification accuracy is retained and the reduction in computational time is lesser than that of dataset I as there are more number of feature vectors with high confidence in the optimal set.

4 Conclusions

We addressed the problem of dimensionality reduction for classification. We proposed a decision-based framework for dimensionality reduction using confidence factor as an evaluation measure for generating a relevant feature subset for specific target. Confidence factor is generated for all features competent for classification using evidence parameters. Evidence parameters are computed based on intersection of classes in the distribution of feature vectors and distance between peaks of distribution in feature vectors and are combined using DSCR. We demonstrated the results of the proposed framework for sky and ground classification using various datasets. The classification in low-dimension space was performed and the classification accuracy was retained optimizing the computational time.

References

1. R. Archibald and G. Fann. Feature selection and classification of hyperspectral images with support vector machines. *Geoscience and Remote Sensing Letters, IEEE*, 4(4):674–677, 2007.
2. C. Bielza and P. Larrañaga. Discrete bayesian network classifiers: a survey. *ACM Computing Surveys (CSUR)*, 47(1):5, 2014.
3. Dempster and A. P. A generalization of bayesian inference. *Journal of the Royal Statistical Society*, 30:205–247, 1968.
4. R. Gupta and R. Kapoor. Comparison of graph-based methods for non-linear dimensionality reduction. *International Journal of Signal and Imaging Systems Engineering*, 5(2):101–109, 2012.
5. R. Jensen and Q. Shen. Semantics-preserving dimensionality reduction: rough and fuzzy-rough-based approaches. *Knowledge and Data Engineering, IEEE Transactions on*, 16(12):1457–1471, 2004.
6. R. U. Kay. Fundamentals of the dempster-shafer theory and its applications to system safety and reliability modeling. *RTA 3-4 Special Issue*, December 2007.
7. J. A. Lee and M. Verleysen. *Nonlinear dimensionality reduction*. Springer Science & Business Media, 2007.
8. G. Sasikala, R. Kowsalya, and M. Punithavalli. A comparative study of dimension reduction techniques for content-based image retrieval. *The International journal of Multimedia & Its Applications (IJMA) Vol*, 2:40–47, 2010.
9. V. Shereena and J. M. David. Comparative study of dimensionality reduction techniques using pca and lda for content based image retrieval.
10. V. Shereena and J. M. David. Significance of dimensionality reduction in image processing.
11. C. O. S. Sorzano, J. Vargas, and A. P. Montano. A survey of dimensionality reduction techniques. arXiv preprint arXiv:1403.2877, 2014.
12. R. Tabib, U. Patil, S. Ganihar, N. Trivedi, and U. Mudenagudi. Decision fusion for robust horizon estimation using dempster shafer combination rule. In *Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG), 2013 Fourth National Conference on*, pages 1–4, Dec 2013.
13. C. Thomas and N. Balakrishnan. Modified evidence theory for performance enhancement of intrusion detection systems. *11th International Conference on Information Fusion*, pages 1–8, July 2008.
14. L. Van Der Maaten, E. Postma, and J. Van den Herik. Dimensionality reduction: a comparative. *J Mach Learn Res*, 10:66–71, 2009.
15. L. van der Maaten, E. O. Postma, and H. J. van den Herik. Dimensionality reduction: A comparative review, 2008.
16. L. Wu and Y. Wen. Weed/corn seedling recognition by support vector machine using texture features. *African Journal of Agricultural Research*, 4(9):840–846, 2009.

A Decision Tree-Based Middle Ware Platform for Deploying Fog Computing Services

Mahesh Sunkari and Raghu Kisore Neelisetti

Abstract Cloud computing has become a cost-effective and reliable distributed computing model for the end users to share IT resources from a pool of computational resources based on their demand in real time. Cloud computing offers advantages of rapid provisioning and release of resources with minimal effort. While cloud computing offers advantages of cost-effective access to sophisticated hardware and software, the turn around time is a hindrance because of unreliable physical layer connectivity especially for business located in a remote location. Not all business solutions need cloud services to the same extent and at all times as the local IT resources available may be sufficient to handle a business issue such as performing analytics on the locally stored data. Fog computing and Grid computing helps in achieving this. Fog computing aims to bring computational power to the edge of the network and by doing so reduces the operational cost and execution time at the cost of accuracy of results. However, a middle ware is required to determine whether an information query needs to be executed in the cloud or if it can be executed on a group of computers that are geographically closer to the business. In this paper, we propose and develop an intelligent middle ware platform that is based on decision trees to optimization the execution of any information query.

Keywords Fog computing · Decision trees · Seattle

M. Sunkari (✉)
School of Computer and Information Sciences, University of Hyderabad,
Hyderabad, Telangana, India
e-mail: smahesh@idrbt.ac.in

R.K. Neelisetti
Institute for Development and Research in Banking Technology, Road No: 1,
Castle Hills, Hyderabad, Telangana, India

P.K. Sa et al. (eds.), *Progress in Intelligent Computing Techniques: Theory, Practice, and Applications*, Advances in Intelligent Systems
and Computing 719, DOI 10.1007/978-981-10-3376-6_33

1 Introduction

Cloud service providers offer IT resources to customers like banks on a need basis and charge them accordingly. Using cloud computing, organizations need not manage the sophisticated IT system on their premises. Cloud service providers also offer the storage space for a low cost which facilitates bank to use it to make their data centralized. Banks can access this data with little effort and at no maintenance cost. In case of large organizations, cloud computing is deployed and managed in house as a private cloud. In such cases organizations take advantage of the elasticity feature of cloud computing.

While a centralized management of all IT resources, works for businesses such as e-commerce companies, a better computing paradigm is necessary for banks. Banks have their branches spread across the globe and in case of India, the government regulations expect the banks to distribute their services between urban and rural areas. While cloud computing allows the bank branches to take advantage of complex analytical algorithms (in terms of time and memory complexity) that can provide very accurate results, the bandwidth costs and turn around time to avail the service becomes hindrance in certain situations. Bank branches often do not need highly accurate results, but rather an approximate estimate for the final result, but in a very short time period is necessary. In such cases, banks trading bandwidth costs and accuracy for lower turn around time and lower algorithmic complexity is an acceptable business practice. Further, cloud computing when applied to banking industry throws two challenges. The first is poor internet connectivity in rural branches. Bandwidth issues in rural branches make it challenging to connect to the cloud, run analytics and get the results back in time during business hours. Second, in urban areas while bandwidth is not a major constraint, network up time is a concern. Network failures happen either because of cable cuts or router issues along the path to the bank's cloud. While the power of cloud (in terms of high end algorithms) cannot be brought to each remote branch, the alternative is to switch to grid computing built using computers with in the branch and/or computers located at branches in the same city and switch to cloud only for high end and more precise computation.

2 Fog and Edge Computing

In a cloud based IT strategy, the data from each branch is logged into a central location and a global analytics is executed on the entirety of the data. Banks also use cloud for handling the internal business operations comprising of small amount of data and are not expected to produce results over a shorter duration of time such as knowing the number of transactions being done in a day, dragging the information of a specific customer, estimating the peak hours of branch operation over a couple of days, estimating the performance of a small group of branches within a city, etc. In the traditional deployment and usage of IT resources, the role of a branch is largely restricted to collecting data and reporting it to the cloud. But as bank branches are distributed all over the globe and that the overall performance of bank can be known only by having the knowledge of individual performance of each branch, it might be better to build a computing platform where in branches are allowed to perform some computation. This would allow branches from overcoming bandwidth limitation in Sect. 1. New computing paradigms in the form of fog and edge computing have been proposed. The idea is to use computational power that is located closer to the point of data generation rather than always relying on cloud. For example, in case of bank branches, the computers have significant processing power and memory resources, they remain under utilized at present as all processing happens on the cloud. The computational power in the branch can be used to process and analyze various internal business operations, day-to-day operations, branch specific and also the group of branches operations to further reduce IT costs and improve redundancy. While cloud computing achieves cost effectiveness through sharing of computational resources and ease of dynamic management through elasticity, fog computing aims to provide an additional feature of quality of resource (QoS). The goal of fog computing is to bring computing closer to data, and thus reduce the turn around time or rather latency associated with analyzing data in case of cloud computing. While fog computing might experience in the accuracy of results, it does so at the cost of redundancy and client objectives. Unlike cloud computing that aims for resource pooling at a centralized geographical location, fog computing aims for local resource pooling between computing clusters spread over a small geographical location.

In case of cloud computing, the physical layer connectivity between the cloud server and the bank is also unreliable and not secure guaranteed especially for business located in a remote location. Due to the heavy traffic being generated in the network, the turn around time is also a hindrance. This can be eradicated by bringing the computational power to the edge of the network. It can be achieved by utilizing the local resources since the physical layer connectivity between bank branches is more stable and reliable.

3 Overview of the Proposed Middle Ware Platform

The attributes of an operation are mainly divided into two characteristics: attributes of data analytics task to be performed and IT resources available. The cost of execution of an operation depends on size of the data set to be analyzed, complexity of the analytics algorithm to be used, bandwidth costs to move the data to the cloud. The total processing time includes the round trip delay to the cloud and time taken to perform analytics. The attributes of the IT resources are bandwidth, up time of the CPU resources and the amount of CPU time that can be availed for a specific job.

In the current work, we aim to deploy the analytical task to be deployed in one of the following three geographical locations:

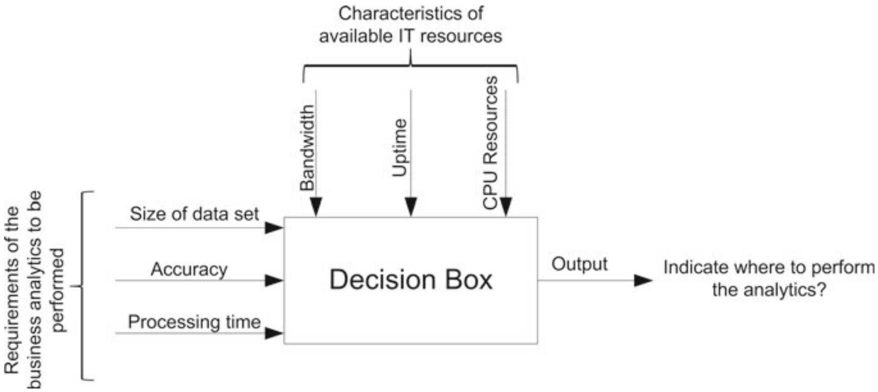

Characteristics of available IT resources
Bandwidth
Uptime
CPU Resources
Requirements of the business analytics to be performed
Size of data set
Accuracy
Processing time
Decision Box
Output
Indicate where to perform the analytics?

Fig. 1 Proposed Model

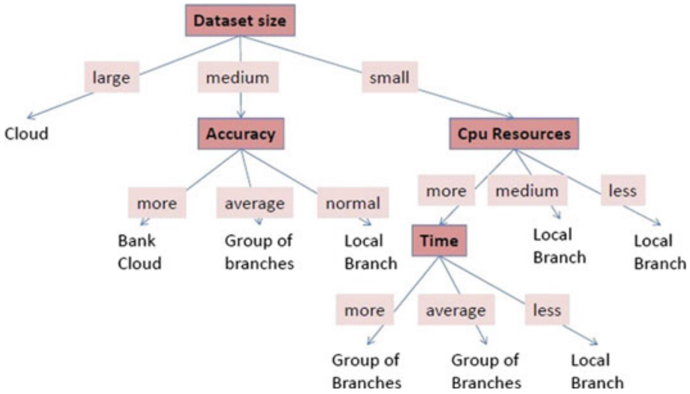


Fig. 2 Classification

- perform on the cloud.
- perform on a group of computers (grouped together to form a grid) located with in the branch.
- perform on a group of computers (grouped together to form a grid) located at different branches of the bank but within the same city or town.

As shown in Fig. 1, a decision box is to be developed such that it acts as a middle ware and decides the appropriate execution place. Whenever a request is made, the decision box looks at the characteristics of IT resources available at that specific time and the requirements of the business analytics to be performed and recommend the best option available.

To build the decision box, in this work we make use of decision trees. Figure 2 shows a hypothetical decision tree for the current problem. The output of the decision tree indicate the appropriate location (one of the 3 locations indicated earlier in the section) for performing business analytics. The aim of the decision tree is to recommend a place that is closer to the source of data for performing analytics.

4 Implementation

4.1 Requirements Gathering

From the observation of bank logs, the analytical characteristic values are estimated and tabulated in Table 1. The data found at a bank is usually less than 40 TB and most of the bank branches are holding the data of size 100 GB or less. In many analytical problem, the branches need quick turn around time and are willing to trade accuracy of the results for turn around time.

The network bandwidth required to handle the various business operations found to be about 40 Mbps. The CPU resources being utilized by the branches is less than 50% of the available resources. The up time of the servers needed to complete the execution of businesss operation is observed to be around 10 h. These values are tabulated in Table 2.

4.2 Construction of Decision Tree

In the current work, we use CART [1] decision tree for classification. CART decision tree is constructed using the attributes of business data to be analyzed and network characteristics. The possible place of execution is used as class variable. The

Table 1 Analytical characteristics

S.No	Dataset size (GB)	Accuracy (%)	Time (hr:mm)	Execution place (Result)
1	100	90	2:00	Data center
2	10	85	0:30	Local branch
3	60	98	4:30	Data center
4	30	89	3:00	Group of branches
5	..	..	..	..
6	..	..	..	..

Table 2 Network characteristics

S.No	N/W bandwidth (Mbps)	Uptime (hr:mm)	CPU resources (%)
1	5	01:10	10
2	12	00:50	25
3	6	00:55	8
4	10	01:28	15
5	..	..	..
6	..	..	..

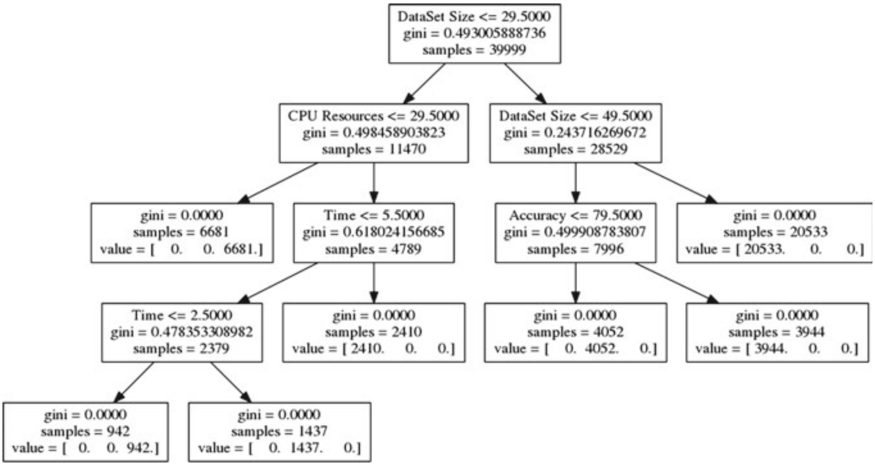


Fig. 3 Decision tree

information needed to construct the initial decision tree is historical data. As the network characteristics keep changing from time to time, we propose using network daemon and ping command to gather data. Using ping command, we can calculate available bandwidth and up time. The daemon runs constantly gathering information about the available CPU resources. The necessary data is gathered for each possible place of execution. Whenever a request for business analytics is made, the network characteristics at that specific moment are measured and along with these values, the analytical characteristics of the request made are passed to the decision tree as input, the decision tree finds the best possible place of execution and then the request is correspondingly redirected. We used scikit-learn [2] to build the CART decision tree. Scikit-learn is a machine-learning tool for data mining and data analysis. The analytical and network characteristics are passed as features and places of execution are passed as class variables. The resulting decision tree is shown in Fig. 3.

In Fig. 3, each node shows the best split condition selected, the impurity measure gini [3] and the number of samples taken into consideration for that specific node characteristics. The size of data set to be processed is taken as the first best split condition to classify the samples into two groups. In left branch, CPU resources is taken as best split condition and in right sub tree, again data set size is selected as best split condition, and so on. At each and every leaf node, the values in matrix indicates the number of samples classified for each execution place such that the left most leaf node infers that number of samples which can be executed in cloud is zero, number of samples which can be executed among a group of branches is zero and which can be executed at local branch are 6681.

4.3 Challenges Overcome

4.3.1 Handling Training Data with Missing Attributes

Two common strategies widely used to deal with missing values are:

- assign the most frequently occurring value in the data set to the missing value.
- Identify all possible values and assign probability to each possible value.

4.3.2 Handling Continuous Attributes

Most widely used technique to deal with continuous attributes is through compartmentalization of the values into several discrete set of intervals. Even though the size of data set to be analyzed can be of any volume, in our study we classify the size of data set into the following three compartments.

- data set of volume 30 GB or less is considered to be small,
- any data set of size more than 30 GB and less than 50 GB is considered to be medium sized,
- any data set size of size 50 GB or more is considered to be large volume.

4.3.3 Finding the Best Split Condition

For the construction of decision tree, we need to identify the best split condition that can produce clear classification of requests. If all the requests are classified into a single final value (execution place) after testing the condition, then it is said to be having zero impurity else if the requests are distributed uniformly for every class, then it is said to be having highest impurity. Statistical metrics such as Entropy, Gini, and Classification error are developed to select the best split. Gini is the measure of impurity and is used to select the best split for our decision tree.

$$gini(t) = 1 - \sum_{i=0}^{c-1} [p(i|t)]^2$$

where,

- $gini(t)$ represents the impurity measure at node t.
- $P(i|t)$ represents the fraction of requests belonging to class i at node t.

- *Calculation of gini impurity measure:*

At root node, entire training data set of 39,999 records is considered. Out of which 26,887 records are labeled as belonging to the *cloud* group, 5489 records belong to the *branches* group and 7623 records belongs to *local* group. Therefore,

$$gini(rootnode) = 1 - ((\frac{26887}{39999})^2 + (\frac{5489}{39999})^2 + (\frac{7623}{39999})^2)$$
$$gini(rootnode) = 1 - 0.5069941$$
$$gini(rootnode) = 0.4930059$$

In subsequent classification, the impurity measure is reduced for at least one branch of tree by choosing the best split condition. It is terminated when there is no significant improvement in gini impurity measure or if the number of records at a node meets a threshold value. A threshold value is the minimum number of records to be present at a node to represent a class and not to be classified further.

4.3.4 Model Over Fitting

There are two types of errors associated with classification model. They are

- Training errors
- Generalization errors

Training errors also known as resubstitution error or apparent error, is the number of mis-classification errors committed on training records, whereas generalization error is the expected error of the model on previously unseen records. An idle model is one that has low training error and low generalization error. It needs to be observed that a model with low training error does not necessarily have low generalization error and in fact may have larger generalization error in case of model over fitting.

4.3.5 Handling Over Fitting in Decision Tree Induction

There are two strategies to handle model over fitting. They are:

- Prepruning:

The algorithm used for building decision tree is stopped based on heuristics so as to avoid perfect classification of training data set. The heuristics prevent the growth of complex sub trees.

- Postpruning:

The major draw back of prepruning technique is the difficulty in identifying the ideal depth of the decision tree even before it has grown and often results in premature termination of the decision tree. This problem is overcome in post pruning techniques by allowing the decision tree to grow to its maximum size before it is pruned. The fully grown tree is then trimmed using a bottom-up approach wherein a subtree is replaced with a leaf node whose class label is determined either by the majority class of records or the most used branch of the sub tree. Further, error estimates can be used to determine the ideal decision tree.

Table 3 Classification of samples

Test Number	Number of samples	Cloud	Group of branches	Local branch	% of samples for cloud
1	10000	7603	445	1952	76.03
2	20000	15636	3699	665	78.18
3	30000	20700	4582	4718	69.00
4	40000	26920	5458	7622	67.30
5	50000	35121	8825	6054	70.24

5 Results

From Fig. 3, it can be observed that the number of samples or operations being classified into cloud are 26,887 out of 39,999 samples and number of samples being classified into group of branches are 5489 and which are classified into local branch group are 7,623. The above experiment has been done several times and the results are tabulated in Table 3.

The average percentage of operations being executed in cloud =

$$\frac{76.03 + 78.18 + 69 + 67.3 + 70.24}{5} = 72.15\%$$

The number of operations being executed in cloud is brought down by nearly $(100 - 72.15) = 27.85\%$ which results in the reduction of cloud service usage.

Performance Evaluation The results of classification algorithm are validated using Holdout method, cross validation, random subsampling, and bootstrap techniques. In Holdout method, the decision tree in Fig. 3 is constructed in such a way that original data is divided into training data set comprising of 80% data and test data set comprising of 20% data. The accuracy of classification was 91%. The results were also evaluated using cross-validation methods.

The classifier's performance was improved by repeating hold out method several times and random subsampling was used to overcome the demerits of holdout method. However, since the process does not utilize as much data as possible for training, it still encounters the problems associated with holdout method as some records might be used for training more often than others. So, we also experimented with bootstrap approach wherein the training records are sampled with replacement.

5.1 Emulation of Proposed Model

Seattle [4] is an open source platform ideally suited for networking and distributed systems research. It is an open, community driven research and educational testbed that offers a large deployment of computational resources in the form of computers, servers, and phones provided by end users across the world. The distributed computing platform composes of end user systems and incorporates essential proper security features such as sand boxing so that the deployed programs operated in a safe and contained manner. In addition the platform allows true owners of the end device to limit the percentage of their computational power that Seattle platform can use.

Seattle provides tools and programming resources to deploy computational algorithms on a desired group of computers. In the current setup, we grouped resources in a the local LAN into one category, computers on the WAN but in the same city (coordinated with resources at another university) and then grouped together a set of VMs on the cloud platform managed by our department. The proposed algorithm was implemented as a middle platform, which activates different group of computers based on the result of the decision tree.

6 Conclusion

In this work, we propose a decision tree-based middle ware platform for taking advantage of fog computing paradigm in the banking environment. The main idea of the proposed model is to overcome the communication challenges faced by banks because of unreliable physical layer connectivity and poor quality of service by moving computation closer to the source data if possible. The cost of using the cloud services is greatly reduced by the efficient use of available computational resources with in the branch and/or at branches with in the same city or town. Based on the network characteristics and availability of computation resources, the proposed decision tree based middle ware platform redirects the computational tasks to different geographical locations. We use the open source Seattle distributed computing platform to emulate the proposed solution. A fall in the cloud usage is also observed from Fig. 3 as a result.

References

1. Loh, Wei-Yin: Classification and regression trees, In: Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, (2011)
2. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine Learning in Python. In: Journal of Machine Learning Research, (2011)

3. Pang-Ning Tan, Steinbach, M., Kumar, V.: Introduction to Data Mining, (First Edition). Addison-Wesley Longman Publishing Co., Inc. (2005)
4. Zhuang, Y., Rafetseder, A., and Cappos, J.: Experience with seattle: A community platform for research and education, In: Research and Educational Experiment Workshop (GREE), Second GENI. IEEE, (2013)

Development of 3D FULL HD Endoscope Capable of Scaling View of the Selected Region

Dhiraj, Priyanka Soni and Jagdish Lal Raheja

Abstract The stereo vision ability of human beings enables the surgeon with depth perception, and thus can allow for organ localization in human body. In this work, we propose an approach of developing a 3D prototype for stereo endoscopes. The stereo calibration and rectification processes have been implemented in order to nullify the effects of lens distortions. The processed output was in interlaced format and can be visualized in 3D format on a passively polarized monitor using polarized glasses. The proposed system also provides the snapshot, video write, and retrieval in individual left and right streams along with live process display. The scaling feature provides a detailed view of the selected region of interest.

Keywords Stereo · Endoscope · Stereo calibration · Image rectification · Depth · Lens distortion · Passive polarization · Full high definition

1 Introduction

The modern advancements in the display devices have modified the facilities in the form of tools in a dynamic and remarkable way for surgical applications. Earlier known techniques such as MRI allows for an approximation view of the patient organs without providing their localization information to the surgeon along with their radiation hazards. The surgeon has to perform the surgery using long incisions and it leads to long recovery times of patients and risks of infection also. To solve this problem, the 2D endoscopes had been used which helps in performing surgery using small incisions but it lacked depth information and also suffers from image

Dhiraj (✉) · P. Soni · J.L. Raheja
CSIR-CEERI, Pilani, India
e-mail: dhirajsangwan@hotmail.com

P. Soni
e-mail: p.soni0109@gmail.com

J.L. Raheja
e-mail: jagdish.raheja.ceeri@gmail.com

P.K. Sa et al. (eds.), *Progress in Intelligent Computing Techniques: Theory, Practice, and Applications*, Advances in Intelligent Systems and Computing 719, DOI 10.1007/978-981-10-3376-6_34

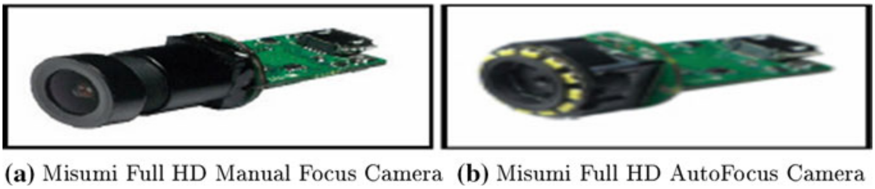

(a) Misumi Full HD Manual Focus Camera (b) Misumi Full HD AutoFocus Camera

Fig. 1 Misumi FULL HD miniature cameras [3]

distortion effects [1]. The lack of depth perception in the observed scene earlier made the endoscopy less effective. The human anatomy requires depth perception by the surgeon in order to make the surgery more effective and result-oriented.

Many researchers had used different techniques for retrieving depth information from the images. In robot assisted surgery, ultrasound images had been used for 3D reconstruction [2]. The computer-aided surgery (CAS) software had also been used for decreasing errors of clinical after effects. The depth information was simulated by using imaging algorithms by regenerating the stereo data from two separate video streams. In this article, a prototype of a real-time FULL HD stereo system using Misumi sensors is shown in Fig. 1 has been demonstrated.

The Misumi MD-B5014-3.0 and Misumi MD-B5014LV-3828 have been used for making the stereo rig for the endoscope. The camera are very small in size with a foot print of 14 × 26 mm only as shown in Fig. 1. In earlier attempts, the 3D was visualized using active shutter glasses by surgeons but it mostly leads to headache, fatigue and nausea [4]. In place of this, the proposed system uses light weight passive polarized glasses with no after effects.

The article consists of six sections. The concept of stereo imaging is described in Sect. 1. The design of stereo system is discussed in Sect. 2. Section 3 covers the logic implementation. Results and analysis are described in Sect. 4. Section 5 contains the conclusion of the research work.

2 Stereo System Design

A prototype of 3D endoscope has been developed with added features in the form of snapshot, video read, write and real-time zoom and unzoom facility. The stereo assembly using the Misumi miniature cameras is shown in Fig. 2.

The stereo system receives the video output of left and right cameras simultaneously. The captured frames were used to perform the stereo calibration [6] and rectification. The rectified right and left images were used to form the combined image having even and odd scan lines from right and left images and this process is called as interlacing. Another version of stereo rig was also designed which has auto focus and 10 white LED’s mounted on the sensor itself to illuminate the view homogeneously as shown in Fig. 3.

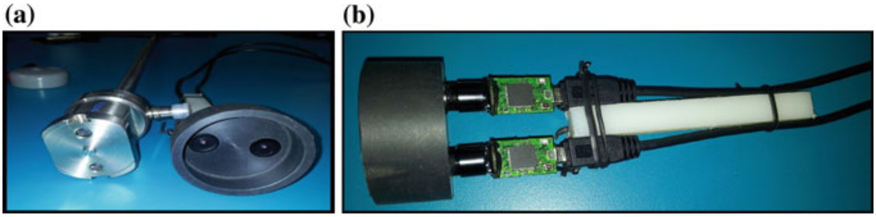


Fig. 2 Stereo endoscope system. **a** Stereo rig with 3D scope [5]. **b** Stereo assembly

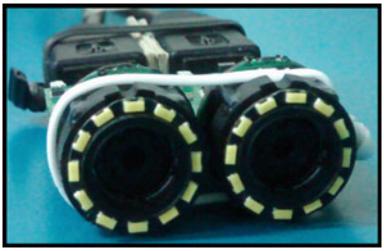


Fig. 3 Misumi MD-B5014LV-3828 stereo assembly with mounted LED

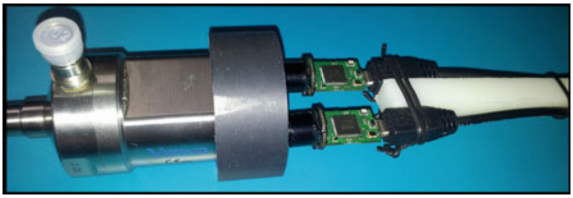


Fig. 4 Stereo endoscope assembly with sensors interfaced

It can capture frames at a maximum resolution of 5.0 M, i.e., 2592 × 1944/15 fps or FHD, i.e., 1920 × 1080 at 30 fps. The final stereo endoscope with FULL HD capture capability is shown in Fig. 4. The Viking Systems dual channel 3D Endoscope was used for the proposed work as shown in Fig. 4. The working length was 415 mm with diameter of 10 mm and field of view as 75°. The Misumi MD-B5014-3.0 as shown in Fig. 4 was used to form a stereo assembly.

The sensors data was captured and transferred to computer using 2 mini USB connections.

3 Logic Implementation

The stereo endoscope produces in contrast to mono scopes [7] two individual streams of video simultaneously from the sensor. To generate the 3D output from them, stereo calibration and rectification needs to be performed on captured data to compensate for the errors and defects in sensors and processes used.

3.1 Stereo Calibration

It is an essential step in stereo vision, so as to generate the metric data from left and right two dimensional images. It was used to compute three types of parameters, i.e., camera intrinsic parameters, camera distortion parameters and extrinsic parameters [8]. A checkerboard pattern is used for this step as it has many identifiable points in the form of corners which were successfully detected [9].

3.2 Stereo Rectification

The stereo rectification was used to reproject the two cameras image planes such that they lie in the forward equivalent formation. First, the images were reprojected and then alignment of the two images in the same line was done [10]. The hartley [11] and bouguet [12] algorithms were tested for stereo rectification and due to the promising results of Bouguet algorithm, it is used for the final prototype implementation.

3.3 Interlaced View Generation

Interlacing is a three dimensional display technique used for converting two frame data into single frame at the cost of loss in resolution [13]. The resultant image as shown in Fig. 5 contains half number of scan lines from left frame and half number of scan lines from right frame.

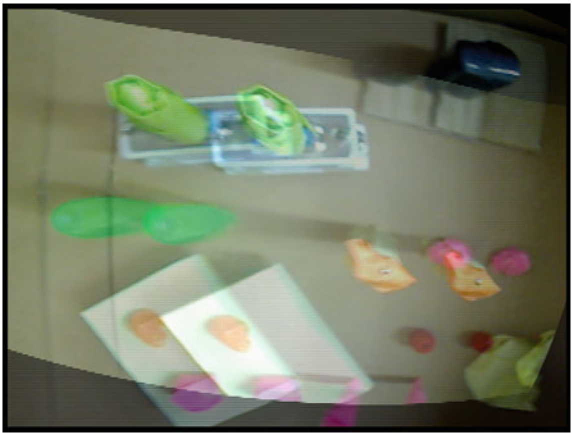


Fig. 5 Interlaced 3D output

This process requires the viewing glasses to be of same polarization as the display screen as shown in Fig. 5. The passively polarization technique was used to project and view the 3D image on viewers eye.

3.4 Zoom View Generation

Zooming is the process which was used, to increase the number of pixels by scaling an image area X of (width*height) data elements by a scaling factor F, so that the image appears larger. The zooming can also be called as scaling of an image. Various methods are reported in literature for zooming like nearest neighbor interpolation [14], bilinear interpolation [15], bicubic interpolation [16], k-times interpolation etc. The illustration of the zooming is given in Fig. 6.

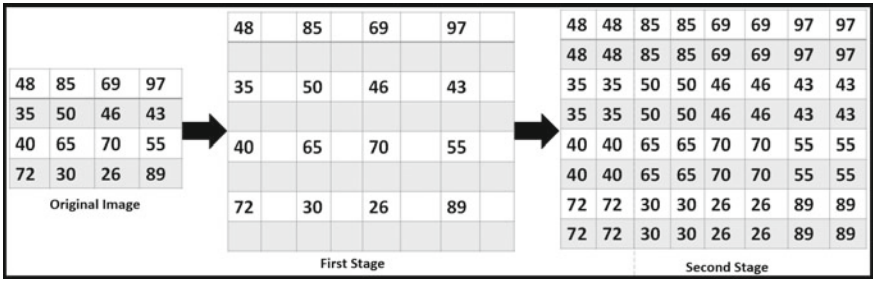
The K-Times interpolation technique was used for zooming process due to its promising results. The flowchart explaining the sequence of steps for the algorithm is given in Fig. 7.

The K-times interpolation technique was used for zooming algorithm. First row wise zooming was done followed by columns wise zooming.

The dimensions of the new image will be as given in Eq. (1).

$$\{K * (number\ of\ rows - 1) + 1\} * \{K * (number\ of\ columns - 1) + 1\} \qquad (1)$$

For a source image of 2 rows and 3 columns as shown in Fig. 8a consider k = 3, i.e., zooming factor is 3. The number of values that should be inserted are equal to $k - 1$, i.e., 3 − 1 = 2. The destination image as obtained after row and column wise zooming is shown in Fig. 8b.



48	85	69	97
35	50	46	43
40	65	70	55
72	30	26	89

Original Image

48		85		69		97	
35		50		46		43	
40		65		70		55	
72		30		26		89	

First Stage

48	48	85	85	69	69	97	97
48	48	85	85	69	69	97	97
35	35	50	50	46	46	43	43
35	35	50	50	46	46	43	43
40	40	65	65	70	70	55	55
40	40	65	65	70	70	55	55
72	72	30	30	26	26	89	89
72	72	30	30	26	26	89	89

Second Stage

Fig. 6 Zoomming process

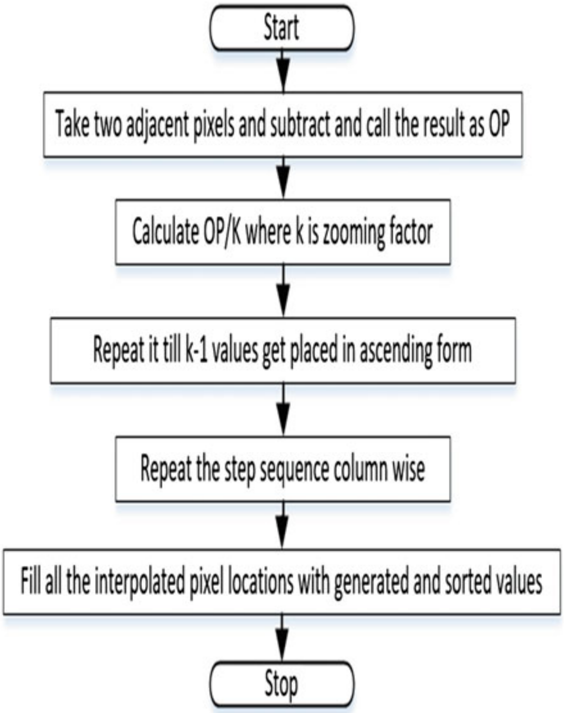
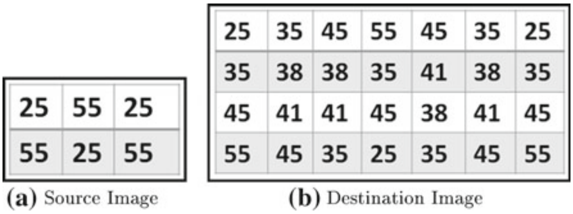

Start
Take two adjacent pixels and subtract and call the result as OP
Calculate OP/K where k is zooming factor
Repeat it till k-1 values get placed in ascending form
Repeat the step sequence column wise
Fill all the interpolated pixel locations with generated and sorted values
Stop

Fig. 7 Flowchart for zoomming technique



25	55	25
55	25	55

(a) Source Image

25	35	45	55	45	35	25
35	38	38	35	41	38	35
45	41	41	45	38	41	45
55	45	35	25	35	45	55

(b) Destination Image

Fig. 8 Source image and destination image

4 Results

The technique provides the flexibility of selecting a region of interest (ROI) as shown in "render" window of Fig. 9, which then acts as input to zooming algorithm. The output of the zooming technique is shown in Fig. 9.

According to the fixed value of "K", the dimensions of the output image were calculated and the intermediate pixel values were evaluated in row followed by column fashion. The algorithm has been successfully implemented using FULL HD Misumi sensors and real-time 3D output has been obtained and displayed on a passive polarized monitor. The algorithm process the left and right video streams by first

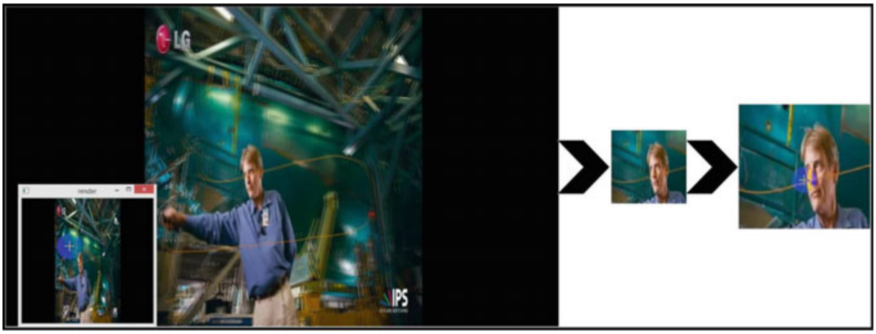


Fig. 9 ZOOM operation on real-time stereo output

performing stereo calibration followed by rectification. The remapped data was then used for interlace output generation which was viewed in 3D form using polarized glasses.

5 Conclusions

The proposed interlaced-based 3D stereo technique has been developed for the robotic assisted surgical applications where a scope is inserted in patient's body and the stereo cameras are used to form a 3D view of internal organs. The proposed method can generate real-time FULLHD output from Misumi miniature cameras. The additional features provided on its API in terms of snapshot capture, 3D video storage, left and right video read and write option, live display of running process, and real-time zooming and unzooming makes the technique more appealing and product-oriented. The passive polarized monitor provides a 3D output using economically priced passive glasses.

Acknowledgements The proposed work was funded by the CSIR-NWP Project, "ASHA". The authors hereby express their acknowledgement to Director, CSIR-CEERI for his motivation and constant support throughout the phase of the work.

References

1. Seth M Brown, Abtin Tabaee, Ameet Singh, Theodore H Schwartz, and Vijay K Anand. Three-dimensional endoscopic sinus surgery: feasibility and technical aspects. *Otolaryngology-Head and Neck Surgery*, 138(3):400–402, 2008.
2. Qinjun Du. Study on medical robot system of minimally invasive surgery. in: Complex medical engineering. In *CME 2007*, pages 76–81. IEEE/ICME, 2007.
3. MISUMI Miniature Camera usb2. http://www.misumi.com.tw. Accessed: 2016-01-24.

4. K Ohuchida, N Eishi, S Ieiri, A Tomohiko, and I Tetsuo. New advances in three dimensional endoscopic surgery. *Journal of Gastroint Digestive Systems*, 3(152), 2013.
5. Viking Endoscope 3d. http://www.conmed.com/products/3dhd-vision-system.php. Accessed: 2016-01-24.
6. Barreto Joao P Melo Rui and Falcao Gabriel. A new solution for camera calibration and real time image distortion correction in medical endoscopy initial technical evaluation. *IEEE Transactions on Biomedical Engineering*, 59(3):634–644, 2012.
7. Ramin Shahidi, Michael R Bax, Calvin R Maurer Jr, Jeremy Johnson, Eric P Wilkinson, Bai Wang, Jay B West, Martin J Citardi, Kim H Manwaring, and Rasoo Khadem. Implementation, calibration and accuracy testing of an image-enhanced endoscopy system. *IEEE Transactions on Medical Imaging*, 21(12):1524–1535, 2002.
8. Adrian Kaehler Gary Bradski. *Learning OpenCV: Computer Vision with the OpenCV Library*. O'Reilly, 2008.
9. Dhiraj Priyanka Soni Jagdish Lal Raheja. Development of 3d endoscope for minimum invasive surgical system. In *2014 International Conference on Signal Propagation and Computer Technology (ICSPCT)*, pages 168–172. IEEE, 2014.
10. Zhongwei TANG Pascal MONASSE, Jean-Michel MOREL. Three-step image rectification. BMVC 2010.
11. Richard I Hartley. Theory and practice of projective rectification. *International Journal of Computer Vision*, 35(2):115–127, 1999.
12. JY Bouguet. The calibration toolbox for matlab, example 5: Stereo rectification algorithm. *code and instructions only*, http://www.vision.caltech.edu/bouguetj/calib_doc/htmls/example5.html.
13. Dhiraj Zeba Khanam Priyanka Soni Jagdish Lal Raheja. Development of 3d high definition endoscope system. In *Information Systems Design and Intelligent Applications*, volume 433 of *Advances in Intelligent Systems and Computing*, pages 181–189. S, February 2016.
14. Rukundo Olivier and Cao Hanqiang. Nearest neighbor value interpolation. *arXiv preprint* arXiv:1211.1768, 2012.
15. Ethan E Danahy, Sos S Agaian, and Karen A Panetta. Algorithms for the resizing of binary and grayscale images using a logical transform. In *Electronic Imaging 2007*, pages 64970Z–64970Z. International Society for Optics and Photonics, 2007.
16. Arthur Sobel and Todd S Sachs. Method of fast bi-cubic interpolation of image information, November 20 2001. US Patent 6,320,593.

An ℓ_1-Norm Based Optimization Approach for Power Line Interference Removal in ECG Signals

Neethu Mohan, S. Sachin Kumar and K.P. Soman

Abstract Accurate analysis and proper interpretation of electrophysical recordings like ECG is a real necessity in medical diagnosis. Presence of artifacts and other noises can corrupt the ECG signals and can lead to an improper disease diagnosis. Power line interferences (PLI) occurring at 50/60 Hz is a major source of noises which could corrupt the ECG signals. This motivates the removal of PLI from ECG signals and is a foremost preprocessing task in ECG signal analysis. In this paper, we deal an ℓ_1 norm based optimization approach for PLI removal in ECG signals. The sparsity inducing property of ℓ_1 norm is used for efficient removal of power noises. The effectiveness of this approach is evaluated on ECG signals corrupted with power line interferences and random noises.

Keywords Power line interferences · ℓ_1 norm optimization · Basis concept · ECG signal analysis

1 Introduction

Recent advances in medical field and research greatly depend on measured electrophysical recordings like electrocardiogram (ECG). For exact disease diagnosis, accurate information extraction from measured biorecordings is required. However,

N. Mohan (✉) · S. Sachin Kumar · K.P. Soman
Center for Computational Engineering and Networking (CEN),
Amrita School of Engineering, Coimbatore, India
e-mail: neethumohan.ndkm@gmail.com
URL: http://www.amrita.edu

S. Sachin Kumar
e-mail: sachinnme@gmail.com

K.P. Soman
e-mail: kp_soman@amrita.edu

N. Mohan · S. Sachin Kumar · K.P. Soman
Amrita Vishwa Vidyapeetham, Amrita University, Coimbatore, India

P.K. Sa et al. (eds.), *Progress in Intelligent Computing Techniques: Theory, Practice, and Applications*, Advances in Intelligent Systems and Computing 719, DOI 10.1007/978-981-10-3376-6_35

these recordings are usually corrupted with different kinds of noises and artifacts, and the presence of which reduces the quality of recorded data by making the analysis a difficult task. Power line interference (PLI) is one of such noise that makes ECG signal analysis a demanding task. PLIs occurring at 50 Hz (Indian scenario) or 60 Hz (European scenario) can corrupt ECG recordings, leading to an improper diagnosis. PLIs are nonstationary in nature and occurs as a thin frequency band near the center frequency. The most challenging task in PLI removal is to remove the interferences without altering the actual ECG signal characteristics [1–3].

Several articles have proposed different techniques for removing PLI in ECG recordings. A typical neural recording system consists of the electrode bundle—a series of amplifiers and filters, followed by digitization and a software signal processing [3]. Several methods are adopted to reduce the effect of interference at the recording hardware level. This bottom-level precautions includes usage of active electrodes, biopotential amplifiers and shielding electrodes in the recording system [1–3]. Despite adopting measures to remove the interference at hardware level, the ECG signals thus recorded will have interferences. The negligible amount of power noises does not corrupt the signal analysis. As the noise level increases, the signal analysis becomes a tedious task. This motivates the use of signal processing algorithms in PLI removal.

The most traditional and classical approach for PLI removal is notch filtering [4]. A notch filter is a band stop filter that attenuates 50/60 Hz PLI frequency. The disadvantage with notch filters is that it causes removal of signal components and introduce signal distortions. Hence even though it is of low cost and computationally efficient it is not widely preferred for PLI removal [2–5]. Several articles published based on adaptive filtering concepts overcome the drawbacks of notch filtering. In adaptive filtering approach, the filter parameters are adjusted according to the variations in the ECG recordings. A comparison between adaptive and nonadaptive filters for interference removal in ECG signals is proposed in [6]. A simplified lattice-based adaptive IIR notch filter [7], a least mean square (LMS) approach [8], a sliding DFT phase locking scheme [9], a State Space Recursive Least Square approach (SSRLS) [10], an extended Kalman filter [11], an H∞ filter [12] are various approaches proposed for PLI removal. Several signal processing algorithms are also have been utilized for PLI removal. In [13], an adaptive filtering approach combined with EMD is used for PLI removal. PLI removal using blind source separation and wavelet analysis combined with EEMD is proposed in [14]. In the previous work [15], the authors have proposed a modified variational mode decomposition (VMD) approach for PLI removal in ECG signals.

In this paper, we are dealing with a sparse regularized optimization approach for PLI removal from ECG recordings. The property of ℓ_1 norm for inducing sparsity on signals is used for PLI and noise removal. The remaining section of the paper is organized as follows: Sect. 2 explains the ℓ_1 norm approach followed by results and discussions in Sect. 3. Finally Sect. 4 concludes the paper.

2 Proposed Approach

We consider the problem of estimating the noise free or clean ECG signal x from the noisy signal y,

$$y = x + p + w \tag{1}$$

where $x, y, p, w \in \mathrm{R}^N$, p is the power line interference, w is the random noises. The sampling frequency is fs Hz. The PLI may vary in amplitude, phase, and frequency [2].

$$p = \sum_{n=1}^{N} a_n \cos(2\pi nft + \phi_n) \tag{2}$$

f is the fundamental frequency, a_n and ϕ_n are the amplitude and phase of the nth harmonics, N is the number of harmonics. The ECG signals are also corrupted with lower order artifacts like electromyogram (EMG) and instrumentation noises. Muscle contraction movements other than heart induces random fluctuations in ECG called EMG noises. Electrical equipments used in ECG device introduce random noises with a white Gaussian distribution and is called as instrumentation noises. To obtain the PLI and random noise-free denoised signal, we follow the ℓ_1 norm-based optimization approach.

2.1 Case 1: General Total Variation Denoising (TVD)

The general formulation of TVD problem [16] is given as,

$$\arg\min_x \left\{ \|y - x\|_2^2 + \lambda \|D_2 x\|_1 \right\} \tag{3}$$

where λ is the regularization parameter and D_2 is the second order difference matrix of size $(N - 2) \times N$ defined as,

$$D_2 = \begin{bmatrix} 1 & -2 & 1 & & & \\ & 1 & -2 & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & 1 & -2 & 1 \end{bmatrix} \tag{4}$$

In this formulation, the ℓ_1 norm of the second order difference of the variables are minimized [16]. The drawback of this general TVD method for PLI removal is the loss of ECG signal components.

2.2 Case 2: Proposed ℓ_1 Norm Denoising

Follows from general TVD and basis pursuit approaches [16, 17], an ℓ_1 norm-based optimization approach is proposed for PLI removal in ECG signals. The formulation of proposed approach is,

$$\arg\min_{x} \left\{ \|y - x\|_2^2 + \lambda_1 \|Ax\|_1 + \lambda_2 \left\|D_2 x\right\|_1 \right\} \tag{5}$$

Here, y is the noisy ECG signal with PLI and x is the denoised signal. In Eq. (5), the regularizers are combination of two simple regularizers. The first term $\|y - x\|_2^2$ ensures that the denoised signal should be close to original signal. However, this term alone will produce a noisy signal. To avoid this, we introduced the second term, $\|Ax\|_1$, which encourages sparsity. A is an orthogonal basis matrix form by sampling sine and cosine waves in the frequency range close to 50/60 Hz. Minimizing ℓ_1 norm of projection of x onto A will ensure the signal x doesn't contain any PLI components. The third term, $\left\|D_2 x\right\|_1$ is the ℓ_1 norm penalty on the second order difference of the variables. D_2 matrix is defined as in Eq. (4). This will reduce the lower order artifacts present in the signal. By applying variable splitting, the problem defined in (5) can be rewritten as,

$$\begin{aligned} &\arg\min_{x,u,v} \left\{ \|y - x\|_2^2 + \lambda_1 \|u\|_1 + \lambda_2 \|v\|_1 \right\} \\ &s.t \qquad u - Ax = 0, \;\; v - D_2 x = 0 \end{aligned} \tag{6}$$

The augmented Lagrangian formulation can be written as,

$$\begin{aligned} L(x, u, v, c, d) = \|y - x\|_2^2 + \lambda_1 \|u\|_1 + \lambda_2 \|v\|_1 + \\ \tfrac{\mu_1}{2} \|u - Ax - c\|_2^2 + \tfrac{\mu_2}{2} \left\|v - D_2 x - d\right\|_2^2 \end{aligned} \tag{7}$$

where $\mu_1, \mu_2 > 0$ are parameters.

In Eq. (5) for PLI removal, an orthogonal basis matrix A is created by sampling sines and cosines waves to capture the PLIs fundamental frequency and its harmonics [18]. The noisy signal is projected onto the basis matrix, A, and by minimizing the ℓ_1 norm of the resultant signal, $\|Ax\|_1$, the interferences can be captured. Since the PLI component usually occurs at the frequencies close to 50/60 Hz and its multiples, bases are created in this close range. To find the correct range at which the bases need to be created, a simple frame-based estimation procedure is adopted. Initially, the average power of entire signal is calculated and the power spectrum is obtained. From this power spectrum the frequency range at which the PLI occurred can be identified. To find this small range, a frame of length 3 is considered in the frequency range 45–65 Hz. The main advantage of choosing this frequency range is that it can accommodate the worst case power line frequency variations [19]. From the frame, the average power is calculated. Then the frame is shifted by 1 Hz (shift = 1 unit), and like this average power of 21 frames (with frame length = 3, shift = 1, 21 frames can be made in the frequency range 45–65 Hz) are calculated. The frame that contains the power noise will have higher average power comparing with other frames. By identifying the frame, the exact frequency range at which PLI occur can be identified and the basis matrix A is created at this frequency range in an interval of 0.5. By minimizing the second-order difference of the variables, $\|D_2x\|_1$, the random noises are reduced. Hence the proposed approach will result in a smoothed ECG signal with reduced noise. The main advantage of the proposed approach is that the ECG signal components are highly preserved.

2.3 Case 3: Sparse Derivative Decomposition (SDD) with Higher Derivatives

For comparing the proposed approach, the sparse derivative decomposition (SSD) and denoising method proposed in [20] is also mapped for PLI removal. In their formulation, the noisy signal y is decomposed in into two components given in Eq. (8).

$$\arg\min_{x_1,x_2}\left\{\|y-x_1-x_2\|_2+\lambda_1\|D_1x_1\|_1+\lambda_2\|D_2x_2\|_1\right\} \tag{8}$$

where D_1 and D_2 represents the first and second order difference matrices. We modified equation (8) by introducing higher order difference matrices D_3 and D_4. The formulation is given as,

$$\arg\min_{x_1,x_2,x_3}\left\{\|y-x_1-x_2-x_3\|_2+\lambda_1\|D_2x_1\|_1+\lambda_2\|D_3x_2\|_1+\lambda_3\|D_4x_3\|_1\right\} \tag{9}$$

where D_2, D_3 and D_4 represents the second, third, and fourth-order difference matrices respectively. By incorporating higher order derivatives the characteristics of ECG signals are highly captured. The second-order derivatives capture the piecewise variation in the signal. Third and higher order derivatives capture the polynomial variations in the signal. The denoised signal is obtained by the addition of these components. All the cases are realized using CVX, a MATLAB-based modelling system for convex programming [21].

3 Results and Discussions

The performance of each case is evaluated on ECG signals under various noise conditions [22]. The ECG data (sampled at 360 Hz) synthetically corrupted with PLI of 60 Hz and its first harmonics is used for performance evaluation. For processing the entire ECG signal, a frame length of 3600 samples (10 second signal) is considered. Figure 1a is the power spectrum density (PSD) of noisy ECG, where the lobs at 60 and 120 Hz indicates the ECG is corrupted with PLI. The performance is evaluated based on noise reduction rate and is expressed in terms of input and output signal to noise ratio (SNR). The input SNR (SNR$_{in}$) is the ratio of the power of the clean ECG signal to the power of the interference and output SNR (SNR$_{out}$) is the ratio of the power of the estimated output signal to the power of the error in the estimation. The strength of PLI is inversely proportional to the SNR value. A lower SNR$_{in}$ indicates that the signal contains high amount of noises. The ECG signal whose SNR$_{in}$ ranges from 30 to −5 db is taken for analysis. The performance of each case on ECG signals corrupted with PLI and random noises also evaluated.

3.1 Case1: TVD with Second-Order Derivative Matrix

The problem defined in Eq. (3) is solved using CVX. Figure 1b is the PSD plot of denoised ECG. The SNR$_{out}$ obtained for case 1 are tabulated in Table 1 as case1. The regularization parameter λ is fixed experimentally based on SNR improvement. For SNR$_{in}$ of 30 to −5 db, λ is fixed between 100 and 5000 range. As the noise level increases, λ has also increases in Eq. (3) to ensure denoising. The performance of this approach is also tested on ECG signals corrupted with EMG and instrumentation noises along with PLI. These noises are additive white Gaussian noise with zero mean and is randomly distributed. Figure 2a shows the noisy ECG sequence and Fig. 2b shows the effect of general TVD for noise reduction. The results of SNR evaluation is given in Table 1 and λ is fixed in 100–5000 range.

3.2 Case2: Proposed ℓ_1 Norm Denoising

The proposed ℓ_1 norm denoising approach with combined regularizers defined in Eq. (5) is applied on ECG data synthetically corrupted with PLI. Initially, the average power and power spectrum of the signal is estimated and a frame of length 3 is defined in 55–65 Hz frequency range. Each frame is shifted by 1 Hz and its average power is calculated. The frame that contains noise will have higher power than other bands (assuming the noise is of high power) and is identified. Since our signal is contaminated with power noises of 60 Hz, the frame 59–61 shows the highest power. Also to identify the frame of harmonics, the same power estimation is done in 115–125 Hz frequency range. The highest power is estimated in a frame of 119–120 Hz indicates the presence of first harmonics at this range. Thus the basis matrix A is created by sampling sines and cosines at both fundamental and harmonics frequency range in an interval of 0.5. The SNR$_{out}$ obtained are tabulated in Table 1 as case 2. The PSD plot of denoised signal is given in Fig. 1c. From the table, it is clear that for all ranges of SNR$_{in}$, proposed approach gives good performance in terms of high SNR$_{out}$. This will conclude that proposed approach is highly insensitive to noise variations. Also on comparison with case 1 (TVD approach), the proposed approach outperforms in terms of high SNR$_{out}$. The signal components are highly preserved in

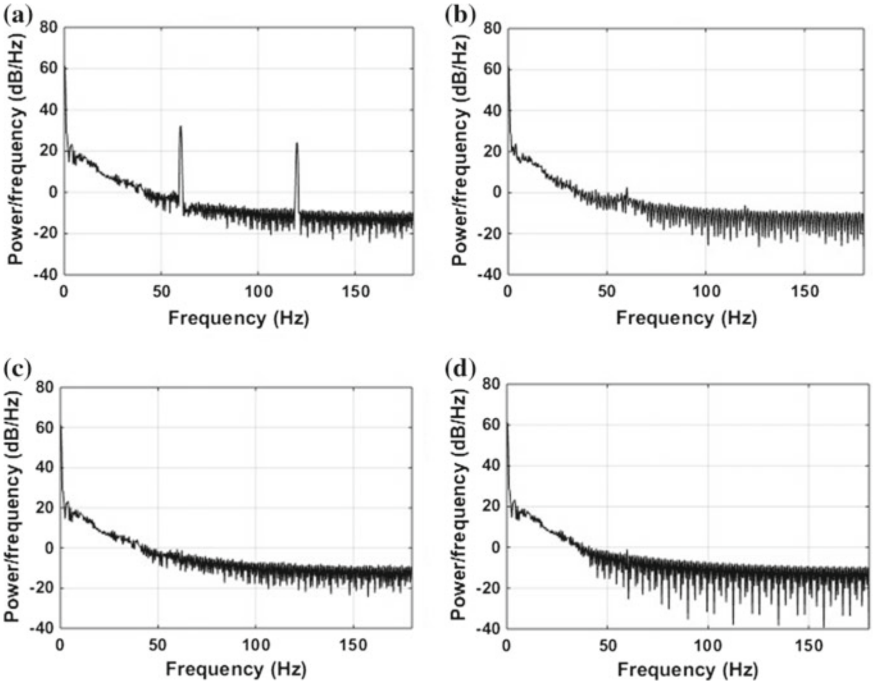


Fig. 1 PSD of **a** noisy ECG signal, **b** denoised signal using case 1, **c** denoised signal using case 2 (proposed approach), where the signal characteristics are highly preserved, **d** denoised signal using case 3, where the signal components are highly lost

case 2 than case 1 and hence it is more appropriate for PLI removal in ECG signals. Also the proposed approach gives good performance even the PLI fundamental frequency is deviated from 60 Hz and which induces higher deviations in harmonics. Under this situation the basis matrix is created by increasing the interval and a high SNR$_{out}$ is achieved.

Based on noise reduction, λ_1 is fixed experimentally. For SNR$_{in}$ of 30–10 db, SNR is improved in 100–1000 range of λ_1. From 10 to −5 db of SNR$_{in}$, good SNR$_{out}$ has achieved in 1000–4000 range. λ_2 is fixed between 0.01 and 0.5 in Eq. (5). Due to the orthogonality nature of basis, projecting the entire signal onto the basis will result in a highly sparse representation (will not capture the required information). Therefore, to capture the required information, the signal is divided into 10 second frames without overlapping.

Now consider the second scenario where ECG signals corrupted with random noises along with PLI. For the PLI reduction the basis matrix is created based on the estimation procedure explained in Sect. 2. The regularization parameters λ_1 and λ_2 are fixed in 100–5000, 100–250 range respectively. Figure 2c shows the effect of proposed approach for noise reduction in ECG signals. The results of SNR evaluation are given in Table 1. The noise reduction without compromising the actual signal

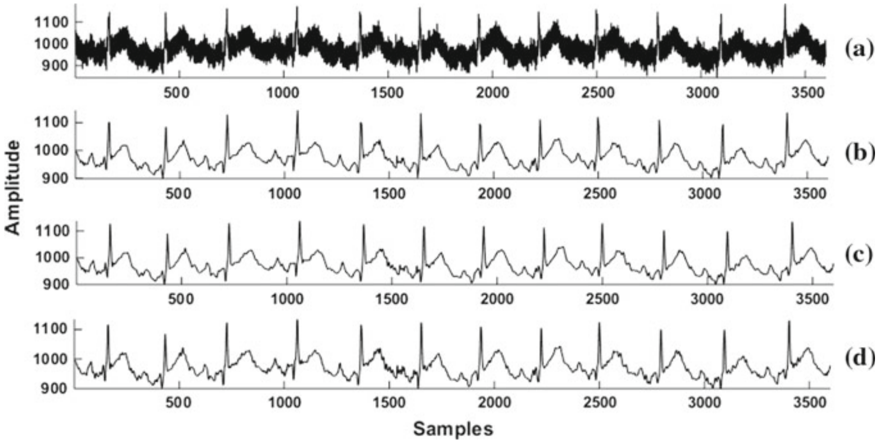


Fig. 2 **a** ECG signal corrupted with random noises and PLI, **b** ECG signal after applying case 1, **c** ECG signal after applying case 2, **d** ECG signal after applying case 3

Table 1 SNR evaluation of three cases; case 1—General TVD, case 2—Proposed ℓ_1 norm denoising, case 3—SDD with higher order derivatives

SNR_{in} (db)	SNR_{out} (db)			SNR_{out} (db)		
	ECG with PLI			ECG with PLI and random noises		
	Case 1	Case 2	Case 3	Case 1	Case 2	Case 3
30.66	48.21	57.54	49.11	44.86	45.78	45.26
22.12	47.13	57.54	45.56	43.49	45.73	45.13
14.17	46.71	57.54	45.81	44.01	45.30	44.48
2.12	45.52	57.57	44.09	43.69	45.48	43.22
0.19	45.51	57.60	43.71	43.69	45.48	42.99
−3.89	44.88	57.59	42.97	42.96	44.43	42.58
−5.82	44.64	57.57	42.35	42.58	43.98	42.25

components is indicated in terms of high SNR_{out}. Comparing with TVD, proposed approach gives high noise reduction in the presence of PLI and random noises.

3.3 Case 3: SDD with Higher Order Derivatives

The proposed approach is compared with SDD approach defined in Eq. (9). Figure 1d shows the denoised signal of SDD approach, where the signal components are highly affected. Results of SNR evaluation are tabulated in Table 1 as case 3. From table it is clear that the noise reduction rate is lesser than case 2. The performance of case 3 under random noise condition is depicted in Fig. 2d.

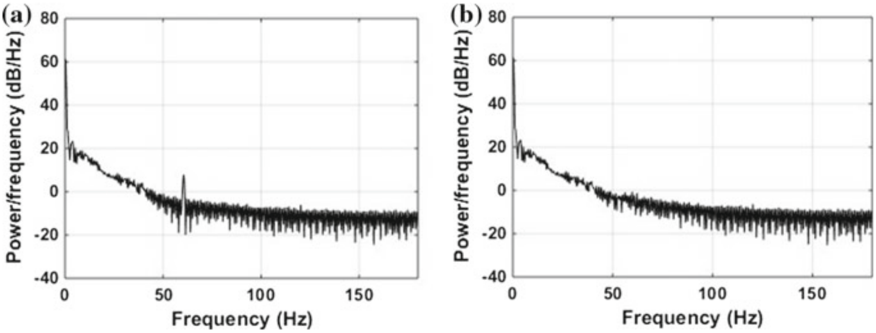


Fig. 3 **a** PSD after applying notch filter, **b** PSD after applying proposed approach

3.4 Comparison with Notch Filtering

The performance of proposed ℓ_1 norm denoising (case 2) is compared with notch filtering approach. Figure 3 shows the effect of a notch filter on a synthetically corrupted ECG sequence. Here the PLI fundamental frequency is slightly deviated from 60 Hz. The notch filter is designed with Q-factor of 35. A high Q-factor will result in a narrowband notch filter. When the PLI frequency is slightly deviated from the fundamental frequency, notch filters fails to capture this frequency deviations and the signal PSD got distorted (Fig. 3a). As seen in Fig. 3b, the proposed approach effectively removes the interference. SNR_{out} is 43.59 in case of notch and 57.52 in proposed ℓ_1 norm denoising approach.

4 Conclusion

This paper deals with an ℓ_1 norm-based optimization approach for noise reduction (PLI and random noises) in ECG signals. The proposed ℓ_1 norm denoising approach is formulated by combining regularizers. This approach evokes a very sharp notch filtering effect to remove the PLI without any external reference signal. Hence the memory requirement is less and is computationally efficient. The ℓ_1 norm of the second-order difference of variables will help to reduce the lower order artifacts present in the signal. The performance of the proposed approach is compared with two cases, case 1—general TVD and case 2—SDD with higher order derivatives. The performance is quantitatively evaluated in terms of input and output SNR on synthetically corrupted ECG signals. Based on evaluation, it is concluded that proposed ℓ_1 norm denoising approach is appropriate for PLI removal in ECG signals.

References

1. Chimene, MF and Pallàs-Areny, Ramon: A comprehensive model for power line interference in biopotential measurements. IEEE Transactions on Instrumentation and Measurement. 49, 535–540 (2000)
2. Keshtkaran, Mohammad Reza and Yang, Zhi: A fast, robust algorithm for power line interference cancellation in neural recording. Journal of neural engineering, 11, 026017 (2014)
3. Thorp, Christopher K and Steinmetz, Peter N: Interference and noise in human intracranial microwire recordings. IEEE Transactions on Biomedical Engineering. 56, 30–36 (2009)
4. Ahlstrom, ML and Tompkins, WJ: Digital filters for real-time ECG signal processing using microprocessors. IEEE Transactions on Biomedical Engineering. 9, 708–713 (1985)
5. Wang, Zheng and Roe, Anna W: Trial-to-trial noise cancellation of cortical field potentials in awake macaques by autoregression model with exogenous input (ARX). Journal of neuroscience methods. 194, 266–273 (2011)
6. Hamilton, Patrick S: A comparison of adaptive and nonadaptive filters for reduction of power line interference in the ECG. IEEE Transactions on Biomedical Engineering. 43, 105–109 (1996)
7. Dhillon, Santpal Singh and Chakrabarti, Saswat: Power line interference removal from electrocardiogram using a simplified lattice based adaptive IIR notch filter. In: 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, vol. 4, pp. 3407–3412 (2001)
8. Bharath, HN and Prabhu, KMM: A new LMS based adaptive interference canceller for ECG power line removal. In: 2012 International Conference on Biomedical Engineering (ICoBE), pp. 68–73 (2012)
9. Mishra, Shivakant and Das, Divya and Kumar, Ravindra and Sumathi, Parasuraman: A Power-Line Interference Canceler Based on Sliding DFT Phase Locking Scheme for ECG Signals. IEEE Transactions on Instrumentation and Measurement. 64, 132–142 (2015)
10. Butt, Muhammad and Razzaq, Nauman and Sadiq, Ismail and Salman, Molly and Zaidi, Tahir: Power line interference removal from ECG signal using SSRLS algorithm.In: IEEE 9th International Colloquium on Signal Processing and its Applications (CSPA), pp. 95–98 (2013)
11. Avendano-Valencia, LD and Avendano, LE and Ferrero, JM and Castellanos-Dominguez, G: Improvement of an extended Kalman filter power line interference suppressor for ECG signals. IEEE Computers in Cardiology. 553–556 (2007)
12. Li, Guo-jun and Zhang, Shu-ting and Hao, Lan-Xiao Xiao-jie and Zhou, Xiao-na: Robust power line interference suppression for ECG signal based on H∞ filter. In: IEEE International Symposium on Bioelectronics and Bioinformatics (ISBB), pp. 143–146 (2011)
13. Zhidong, Zhao and Chan, Ma: A novel cancellation method of powerline interference in ECG signal based on EMD and adaptive filter. In: 11th IEEE International Conference on Communication Technology (ICCT 2008), pp. 517–520 (2008)
14. Akwei-Sekyere, Samuel: Powerline noise elimination in biomedical signals via blind source separation and wavelet analysis. PeerJ. 3, 1086 (2015)
15. Mohan, Neethu and Kumar, Sachin and Poornachandran, Prabaharan and Soman, KP: Modified Variational Mode Decomposition for Power Line Interference Removal in ECG Signals. International Journal of Electrical and Computer Engineering (IJECE). 6, (2016)
16. Chambolle, Antonin: An algorithm for total variation minimization and applications. Journal of Mathematical imaging and vision. 20, 89–97 (2004)
17. Chen, Scott Shaobing and Donoho, David L and Saunders, Michael A: Atomic decomposition by basis pursuit. SIAM review. 43, 129–159 (2001)
18. Soman, KP and Ramanathan, R: Digital Signal and Image Processing-The Sparse Way. Isa Publication (2012)
19. Dugan, Roger C and McGranaghan, Mark F and Beaty, H Wayne: Electrical power systems quality. McGraw-Hill, New York (1996)
20. Ning, Xiaoran and Selesnick, Ivan W: ECG enhancement and QRS detection based on sparse derivatives. Biomedical Signal Processing and Control. 8, 713–723 (2013)

21. Grant, M and Boyd, S: CVX: Matlab software for disciplined convex programming, version 2.0 beta, Sep (2012). http://cvxr.com/cvx
22. MIT-BIH Arrhythmia database, https://www.physionet.org/cgi-bin/atm/ATM

Exploration of Many-Objective Feature Selection for Recognition of Motor Imagery Tasks

Monalisa Pal and Sanghamitra Bandyopadhyay

Abstract Brain–Computer Interfacing helps in creation of a communication pathway between brain and external device such that the biological modality of performing the task could be bypassed. This necessitates fast and reliable decoding of brain signals which mandate feature selection to play a crucial role. The literature discloses the improvement in performance of left/right motor imagery signal classification with many-objective feature selection where several classification performance metrics have been maximized for obtaining a good quality feature set. This work analyses the classification performance by varying the feature dimension and number of objectives. A recent many-objective optimization coupled with objective reduction algorithm viz. α-DEMO has been used for modeling the feature selection as an optimization problem with six objectives. The results obtained in this work have been statistically validated by Friedman Test.

Keywords Brain—computer interfacing · Electroencephalography · Feature selection · Many-objective optimization · Friedman test

1 Introduction

Artificial means of decoding and encoding brain signals to assist motor-disabled people have been one of the key research areas of Brain—Computer Interfacing (BCI) since decades. Among several other methods, Electroencephalography (EEG) is often the choice for brain signal acquisition because of its high temporal resolution, non-invasiveness, affordability, and portability [1, 2]. However, for efficient real-time BCI applications fast feature estimation as well as fast classification with sufficient accuracy are the major requirements [1, 2]. Feature selection methods not

M. Pal (✉) · S. Bandyopadhyay
Machine Intelligence Unit, Indian Statistical Institute, Kolkata 700108, India
e-mail: monalisap90@gmail.com

S. Bandyopadhyay
e-mail: sanghami@isical.ac.in

P.K. Sa et al. (eds.), *Progress in Intelligent Computing Techniques: Theory, Practice, and Applications*, Advances in Intelligent Systems and Computing 719, DOI 10.1007/978-981-10-3376-6_36

only helps in meeting these challenges but also helps to reduce the number of redundant and irrelevant features thereby making the classification technique computationally less expensive.

There are several works on feature selection for motor imagery EEG signal classification which can be found in the literature review portions of [1, 2]. Some of these works consider filter approach where the information content of the feature set is maximized using some information theoretic measure and the rest of them consider wrapper approach where the fitness of the feature set is judged based on the classification performance. Most of the previous approaches consider maximizing classification accuracy as the only objective for feature selection [1]. A few of the works which considers multi-objective feature selection are mainly bi-objective formulations where either precision along with recall are maximized [3] or rate of classification error along with the cardinality of feature set are minimized [4]. The work in [2] shows that feature selection using maximization of several classification performance metrics improves the performance of left/right motor imagery signal classification. However, the work does not study the variation of the performance with different feature dimension and whether all objectives are relevant.

In this work, the same dataset, feature extraction procedure, and classifier are used as those in [2]. It considers the claims in [2] to be true. Specifically, the classification accuracy improves from the approach where feature selection is not used to the approach where many-objective feature selection is used. Another observation made in [2] is the reduction in classification time due to smaller feature dimension of the test samples. Less classification time has high importance in real-time motor imagery classification work. This work applies a differential evolution based many-objective optimization (MaOO) viz. α-DEMO [5] and investigates the performance by varying the feature dimension and the number of objectives. Following this, the proposed approach is compared with a previous work [2] and with another popular Genetic Algorithm based MaOO algorithm. The results are statistically validated to conclude the significance of the proposed α-DEMO based feature selection strategy.

In Sect. 2, the overview of the experiment is briefly described. Section 3 presents and discusses the results obtained in course of the work. Finally, the conclusion is drawn in Sect. 4 providing direction for future work.

2 Experimental Paradigm

A typical BCI system consists of a forward non-neuromuscular path connecting the subject's brain to a computer (might be a wired or a wireless connection) and a feedback path through which an external (rehabilitative) device is operated based on computer-generated control signals. The operation of the external device serves as a stimulus to which the subject's brain responds. In this section, the outline of the experiment (forward path) is presented.

Table 1 Different stages of BCI system

Stages	Details
Data acquisition	BCI Competition 2008—Graz Dataset B (250 Hz left/right motor imagery EEG from nine subjects with two sessions/subject)
Preprocessing	Band-pass filtering with passband 0.5–100 Hz, Electrooculogram (EOG) artifact removal using regression between EOG and EEG
Feature extraction	One-sided Power Spectral Density Estimation by Welch's Method (EEG from C3, C4, and Cz electrodes, feature dimension = 378)
Classification	Linear Support Vector Machine (LSVM)

2.1 Major Building Blocks of the BCI System

The forward path of the BCI system consists of the following building blocks: brain signal acquisition, signal preprocessing, feature extraction, feature selection (during training phase), and classification. The specifications of the dataset and other stages except feature selection are summarized in Table 1 and the feature selection stage is explained in detail, in the next section. For more details, the work in [2] and its references can be consulted.

2.2 Many-Objective Feature Selection

Feature selection is the task of choosing a subset of the complete feature set such that the performance is not compromised. The selected feature set has a size $d(<D)$ where D is the size of complete feature set ($D = 378$, here).

For formulating the task of feature selection as a MaOO problem, the encoding of the different candidates of the population, the several objectives and the stopping condition are to be stated.

A population candidate ($\boldsymbol{X}$) is defined as $\boldsymbol{X} = \{x_1, x_2, \ldots, x_d\}$ where each $x_i \in \{1, 2, ..., D\}$ for $i = 1, \ldots, d$. A candidate represents a solution of feature selection. Here, the encoding implies that d out of D features are to be selected and x_i is the index of i-th choosen feature. This work considers feature selection in a wrapper approach. The dataset in terms of the selected features are classified using LSVM and the classification performance indicates the quality of the choosen feature set. Maximization of classification performance metrics [2] viz. Precision, Recall, Accuracy, F1-score, Specificity, and Cohen's Kappa Coefficient are the six objectives considered for this work. The optimizer is stopped when maximum number of generations is attained. A recent differential evolution based optimization algorithm [5] viz.

α-DEMO is used, which is not only a many-objective optimizer but also performs objective reduction for the provided value of α.

3 Analysis of Experimental Outcomes

The experiment is executed on MATLAB R2012b on a computer having 64-bit Core i3 processor @ 2.3 GHz and 4GB RAM. All the results presented, here, are the average value of 50 runs of the proposed approach.

Selection of the objectives out of the six objectives based on the value of α are presented in Table 2. These observations imply that any two out of the three metrics (Precision, Recall and F1-score) are sufficient to express the information given by all of them (row 3 and 4). As Cohen's Kappa assesses the observed accuracy with respect to the expected accuracy, it is of more importance than accuracy itself (row 1, 2 and 5). However, when the complete confusion matrix can be estimated from the given information, observed accuracy suffices (row 4).

Following this, the performance of the MaOO feature selection is studied by varying the dimension of reduced feature set (d). Due to lack of space, the variation in the performance with d is shown only for one of the metrics and for the first session of data acquisition. The variation of Kappa coefficient (as it is best metric from row 1 of Table 2) with d is given in Fig. 1 for all the subjects. The graphs indicate that the best performance occurs when $d = 5$ and is poorer for both lower as well as higher values of d. This observation is also validated using principal component analysis [6] on the complete feature set which reveals that highest five eigenvalues out of the 378 values contains about 90% of the information. The selected features are the spectral estimates between 8–12 Hz (μ band) which contains informative motor imagery signals [7]. Another thing to note is that the best learning takes place for subject 4 followed by subject 5, which is in agreement with the observation made in [2] which uses the same dataset for classification.

For justifying the choice of MaOO algorithm, the feature selection is modelled using two other MaOO algorithms viz. DEMO [2] and Non-dominated Sorting Genetic Algorithm—II (NSGA-II) [8]. The parameter selection of these algorithms

Table 2 Choice of objectives

α	No. of objectives	Recall	Precision	Accuracy	F1-score	Specificity	Kappa
0.1667	1						✓
0.3333	2				✓		✓
0.5000	3	✓	✓			✓	
0.6667	4	✓		✓	✓	✓	
0.8333	5	✓	✓		✓	✓	✓

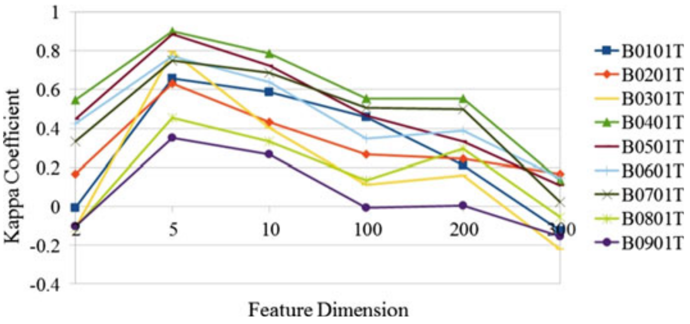


Fig. 1 Variation of performance with different feature dimension (lower dimensions are zoomed)

Table 3 Parameters of different evolutionary algorithms

Algorithms	Parameters and values
DEMO [2]	$F \in [0, 2]$, $CR = 0.8$
α-DEMO [5]	$\alpha \in \{1/6, 2/6, 3/6, 4/6, 5/6\}$, $K_0 = K_1 = 20$, $\beta_1 = \beta_2 = 0.75$
	$F \in [0, 2]$, $CR = 0.8$
NSGA-II [9]	Mutation distribution index = Crossover distribution index = 20
	(Simulated Binary Crossover and Polynomial Mutation)
	Tour size = 2 (Tournament Selection)
Common parameters	Size of Population = 25, Maximum generations = 200

are given in Table 3. The results are statistically validated using Friedman Test [10] which follows a chi-squared distribution with $(k_a - 1)$ degrees of freedom as shown in Eq. (1). The number of datasets (N_S) is 18 considering both the sessions of the nine subjects, the number of approaches used for comparison (k_a) is 3 and the average rank (R_j) of the j-th approach is obtained from Table 4 where the ranks are according to the Kappa values. According to the null hypothesis, all the approaches are equivalent. The algorithms are executed with the four objectives (row 4 of Table 2) because the highest Kappa values are attained for 15 out of 18 datasets while using α-DEMO with $\alpha = 0.6667$.

$$\chi_F^2 = \frac{12N_S}{k_a(k_a + 1)} \left[\sum_j R_j - \frac{k_a(k_a + 1)^2}{4} \right] . \tag{1}$$

Substituting all the values in Eq. (1), the obtained value of $\chi_F^2 = 8.88 > 5.99$ (critical value for 2 degrees of freedom and 95% confidence interval). Thus, the null hypothesis is rejected which validates our claim that the approaches are ranked

Table 4 Ranks of algorithms for Friedman Test

Feature selection using	NSGA-II	DEMO	α-DEMO
Average kappa coefficient	0.6058	0.6180	0.6667
R_j	2.70	1.97	1.15

according to Table 4 where α-DEMO outperforms the others. This is because (i) unlike DEMO and like NSGA-II, α-DEMO is an Ellistist approach [5], and (ii) unlike NSGA-II, α-DEMO limits the number of rank-one solution in the population which reduces risks of trapping in the local optima [5]. Better performance of DEMO over NSGA-II is due to ranking strategy where besides non-dominated sorting, the minimum distance from ideal point is considered [2].

4 Conclusion

The literature reveals that many-objective feature selection is a better approach than single-objective feature selection, and reduction in feature dimension is crucial for real-time application of BCI. This work studies various parameter estimation for performing feature selection using MaOO approach. The feature dimension as well as the number objectives are varied to study the optimal case. Later it is shown that α-DEMO outperforms two popular MaOO algorithms (DEMO and NSGA-II).

Two disadvantages of the work that still prevail are that the feature dimension and the reduced size of objective set have to be user-specified. For further extension of this work, different encoding of the candidates of MaOO might be tried such that the optimal feature dimension is automatically determined. In a similar way, automatic selection of optimal number of objectives (using other MaOO approaches) might be taken up as a future work.

References

1. Pal, M., Bhattacharyya, S., Roy, S., Konar, A., Tibarewala, D.N., Janarthanan, R.: A bacterial foraging optimization and learning automata based feature selection for motor imagery EEG classification. In: International Conference on Signal Processing and Communications (SPCOM), pp. 1–5. IEEE (2014)
2. Pal, M., Bandyopadhyay, S.: Many-objective Feature Selection for Motor Imagery EEG Signals using Differential Evolution and Support Vector Machine. In: International Conference on Microelectronics, Computing and Communication (MicroCom), pp. 1–6. IEEE (2016)
3. Ekbal, A., Saha, S., Garbe, C.S.: Feature selection using multiobjective optimization for named entity recognition. In: 20th International Conference on Pattern Recognition (ICPR), pp. 1–4. IEEE, (2010)

4. Xue, B., Fu, W., Zhang, M.: Multi-objective feature selection in classification: a differential evolution approach. In: Simulated Evolution and Learning, pp. 516–528. Springer International Publishing (2014)
5. Bandyopadhyay, S., Mukherjee, A.: An Algorithm for Many-Objective Optimization with Reduced Objective Computations: A Study in Differential Evolution. IEEE Trans. Evol. Comput. 19(3), 400–413 (2015)
6. Yu, X., Chum, P., Sim, K.B.: Analysis the effect of PCA for feature reduction in non-stationary EEG based motor imagery of BCI system. Optik-International Journal for Light and Electron Optics. 125(3), 1498–1502 (2014)
7. Pfurtscheller, G., Brunner, C., Schlögl, A., Da Silva, F.L.: Mu rhythm (de) synchronization and EEG single-trial classification of different motor imagery tasks. NeuroImage. 31(1), 153–159 (2006)
8. Deb, K., Pratap, A., Agarwal, S., Meyarivan, T.: A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE Trans. Evol. Comput. 6(2), 182–197 (2002)
9. Raghuwanshi, M.M., Kakde, O.G.: Survey on multiobjective evolutionary and real coded genetic algorithms. In: Proceedings of the 8th Asia Pacific Symposium on Intelligent and Evolutionary Systems, pp. 150–161. (2004)
10. Demšar, J.: Statistical comparisons of classifiers over multiple data sets. J. Mach. Learn. Res. 7, 1–30 (2006)

Euler-Time Diagrams: A Set Visualisation Technique Analysed Over Time

Mithileysh Sathiyanarayanan and Mohammad Alsaffar

Abstract Over the years, there is a large amount of data being generated in almost all the fields such as engineering, medical, bioscience and social network. To visualise set relationships for a large data is always a challenge, especially with the data evolving over time. There is no tool that supports the set visualisation represented over time. So, we explored this impossibility by developing a novel visual method and a software tool to generate Euler-time diagrams, which will represent set relations with respect to time. The idea was taken from the well-known visualisations: Euler diagrams and time-series. Euler diagrams represent set relations and time-series represent a sequence of events happened over a time. We merged the idea of Euler diagrams and time-series to spark the novelty. Pattern discovery not only plays an important role in graph analysis but also in the set analysis process. In this paper, we took a case study from the World Health Organisation (WHO) who is constantly trying to understand relationships between various diseases people are affected over a period of time. This motivated us to develop the set relationship time tool, by considering two levels: aggregation and relationships, using data-driven documents (D3) and Google developing tool kit. This prototype tool can be enhanced by considering gestalt principles, topological properties, perceptual and cognitive theories which will help in analysing and interpreting data efficiently.

Keywords Euler diagrams · Knowledge discovery · Social network · Informatics · D3

M. Sathiyanarayanan (✉) · M. Alsaffar
School of Computing, University of Brighton, Brighton, UK
e-mail: s.mithileysh@gmail.com

M. Alsaffar
e-mail: mohammad.alsaffar@gmail.com

P.K. Sa et al. (eds.), *Progress in Intelligent Computing Techniques: Theory, Practice, and Applications*, Advances in Intelligent Systems and Computing 719, DOI 10.1007/978-981-10-3376-6_37

1 Introduction

Data in various fields are evolving over time constantly and our ability to understand and analyse them are quite a challenging task. Understanding a huge set of data manually is far impossible in the current scenario, so a visualisation tool is required to understand set relationships changing over time.

Euler diagrams [1] are a popular tool that represents set relationships and they are used in the field of engineering, medical, bioscience and social network. Euler diagrams are a static representation (as shown in the Fig. 1) which gives useful information about how sets are related. But, they have not been represented dynamically and interactively and there is no other tool to support the set visualisation over time. To visualise set relationships over time for a large data is always a challenge. So, we explored this impossibility by developing a novel visual method and a software tool to generate Euler-time diagrams, which will represent set relations with respect to time. The idea was taken from the well-known visualisations: Euler diagrams and time-series. Euler diagrams represent set relations and time-series represent sequence of events happened over a time. We merged the idea of Euler diagrams and time-series to spark the novelty.

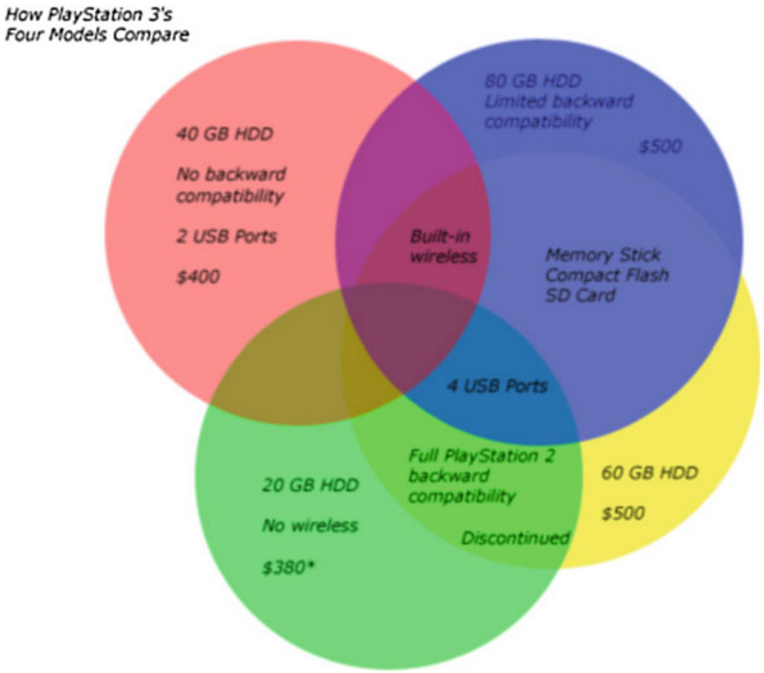


Fig. 1 An example of a static Euler diagram that compares the features on different models of PlayStation 3 gaming consoles. *Source* http://www.wired.com

Since Euler diagrams and Venn diagrams are used in visualising set-based relationships using closed contours. We tried to understand other set visualisations such as Euler-graph diagrams [2, 3], Euler View [4], Untangled Euler diagrams [5], Bubble Sets [6], LineSets [7], KelpFusion [8], linear diagrams [9], treemaps [10, 11] and spherule diagrams [12, 13] but these will introduce more noise which may reduce readability. Pattern discovery not only plays an important role in graph analysis but also in the set analysis process. Due to set complexities, information and elements make the patterns entangled and it becomes difficult for the analysts who are interested in visualising set overlaps over time.

In this article, we took a case study from the World Health Organisation (WHO) who is constantly trying to understand relationships between various diseases people are affected over a period of time. This motivated us to develop the set relationship time tool, by considering two levels: aggregation and relationships, using data-driven documents (D3) and Google developing tool kit using some cooked-up (fake) data. This prototype tool can be enhanced by considering gestalt principles, perceptual and cognitive theories which will help in analysing and interpreting data efficiently.

The rest of the paper is organised as follows: Sect. 2 describes the method we implemented to generate Euler-time diagrams. Section 3 concludes the paper by addressing the limitations and future work.

2 Method

Set representation in a timeline is not a cumbersome process given that the time axis is fixed. When an user or an analyst selects the "period of time" he wants to analyse, he can select the years and the Euler diagrams will be displayed. To visualise a set of data, we considered two levels: aggregation and relationships:
Relationships: Euler diagrams are drawn with a closed contour and mostly with circles where (a) set intersection (crossing) is represented by two circles overlapping (b) set exclusion (disjointness) is represented by two circles not overlapping or touching each other (c) set inclusion (subset) is represented by a circle inside another circle completely. In our tool, Euler diagrams are generated based on the general principles of the set relationship properties. These relationships are coloured and are represented over time in a chronological order (past year to the current year).
Aggregation: Each circle is a visual indicator of a situation over time and the size of the circle represents an aggregated information. Each circle has a label (given in text) in the right side. The Euler diagrams over time are represented in a chronological order (past year to the current year). Also, we considered interaction in our tool. When you mouse over, the information of each set will be displayed.

In the example, we considered a case study from the World Health Organisation (WHO) who is constantly trying to understand relationships between various diseases people are affected over a period of time. Rather than taking a real data (due to many issues), we cooked-up (fake) the data for developing this tool, Figs. 2 and 3.

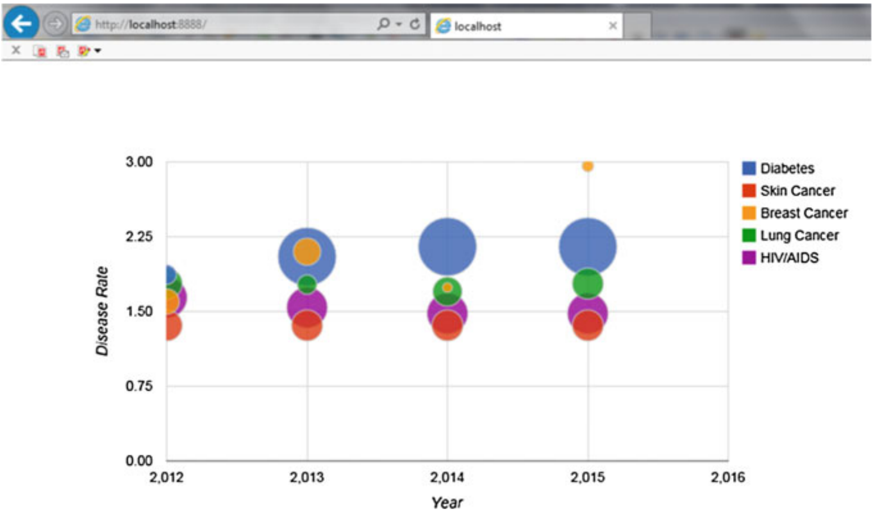


Fig. 2 The screenshot of an Euler-time diagram: a set visualisation technique analysed over time, where the tool is developed using D3 and Google tool kit

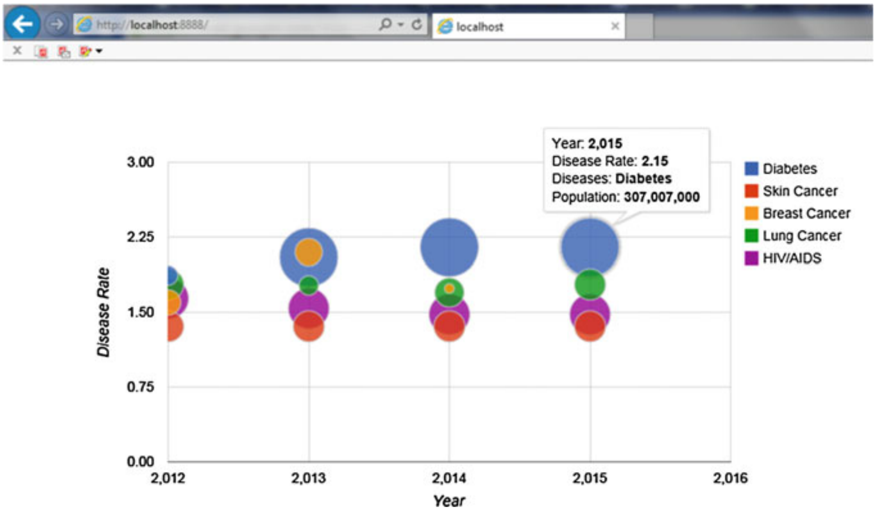

Year: **2,015**
Disease Rate: **2.15**
Diseases: **Diabetes**
Population: **307,007,000**

Fig. 3 The screen shot of an Euler-time diagram demonstrating user interaction

Relationships: The sets represent people affected with various diseases. The overlaps represent some set of people had the diseases in common. For example, in 2013 some people affected with lung cancer had diabetes in common and also some set of people had HIV/AIDS in common, though few people had only lung cancer (no overlaps). Interestingly, all ladies who had breast cancer had diabetes. Now, checking the relationships over time: people affected by HIV/AIDS in 2012 were very low but

over the next few years, it has increased, based on the disease rate and population. In this way, set relationships can be visualised over time.
Aggregation: The circle size for HIV/AIDS has increased due to increased number of affected people. As the number of people increase with respect to time, the size of the circle will also increase and it holds good for decrease of circle size as well.

The tool is an interactive one (see the Fig. 3): when you mouse over, the information of each set will be displayed. For example, when you mouse over to Diabetes in 2015, it gives you the details of the disease rate and population (number of people affected with the disease). This kind of tool will help World Health Organisation (WHO) to promote more public health awareness programmes on a particular disease.

3 Conclusion and Future Work

We have presented a general approach to generate Euler-time diagrams, which represent set relationships over time. This general approach has helped us build a D3 visualisation tool using Google tool kit by considering two levels: aggregation and relationships. The work is at an early stage of development. Our intention is to develop an efficient and effective Euler diagrams with timeline. Sometimes, there is a strong feeling that it is just circles whose position indicates the year and disease rate rather showing Euler diagrams. That is because the circles overlapping are adding noise and there is no colour distinction when sets are overlapped. One of the limitations is that, we used some cooked-up (fake) data rather a real data from a knowledge discovery point of view. As a future work, we aim to use real data and then run user studies to establish any trade-off involved. The other limitation is that, user's attention might get diverted when too much of information is displayed using interaction types. So, we need to try to visualise the core information statically and only if this is not possible (e.g. the problem is too complex), we will consider interactivity to improve readability. Also, this prototype tool can be enhanced by considering gestalt principles, topological properties, perceptual and cognitive theories which will help in analysing and interpreting data efficiently.

Acknowledgements We would like to thank Luana Micallef of Helsinki Institute for Information Technology (HIIT) for her constructive comments.

References

1. M. Sathiyanarayanan and J. Howse, "Well-matchedness in euler diagrams," in *Euler Diagrams (ED), 2014 4th International Workshop on*, vol. 1244. CEUR Workshop Proceedings, July 2014, pp. 16–22.

2. M. Sathiyanarayanan, G. Stapleton, J. Burton, and J. Howse, "Properties of euler diagrams and graphs in combination," in *Visual Languages and Human-Centric Computing (VL/HCC), 2014 IEEE Symposium on*, July 2014, pp. 217–218.
3. M. Sathiyanarayanan, "Drawing euler diagrams and graphs in combination," in *The Theory and Application of Diagrams, 2014 8th International Conference on*. Diagrams Graduate Symposium, July 2014.
4. P. Simonetto, "Visualising of overlapping sets and clusters with Euler diagrams," Ph.D. dissertation, Universite Bordeaux, 2012.
5. N. Riche and T. Dwyer, "Untangling Euler diagrams," *IEEE Transactions on Visualization and Computer Graphics*, vol. 16, no. 6, pp. 1090–1099, 2010.
6. C. Collins, G. Penn, and S. Carpendale, "Bubble sets: Revealing set relations with isocontours over existing visualizations," *IEEE Transactions on Visualization and Computer Graphics*, vol. 15, no. 6, pp. 1009–1016, 2009.
7. B. Alper, N. H. Riche, G. Ramos, and M. Czerwinski, "Design study of linesets, a novel set visualization technique." *IEEE Trans. Vis. Comput. Graph.*, vol. 17, no. 12, pp. 2259–2267, 2011.
8. W. Meulemans, N. Riche, B. Speckmann, B. Alper, and T. Dwyer, "Kelpfusion: A hybrid set visualization technique," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 19, no. 11, pp. 1846–1858, Nov 2013.
9. P. Rodgers, G. Stapleton, and P. Chapman, "Visualizing sets with linear diagrams," *ACM Transactions on Computer-Human Interaction (TOCHI)*, vol. 22, no. 6, p. 27, 2015.
10. M. Sathiyanarayanan and N. Burlutskiy, "Design and evaluation of euler diagram and treemap for social network visualisation," in *Communication Systems and Networks (COMSNETS), 2015 7th International Conference on*, Jan 2015, pp. 1–6.
11. M. Sathiyanarayanan and N. Burlutskkiy, "Visualizing social networks using a treemap overlaid with a graph," in *Computer Vision and the Internet (VisionNet), 2015 2nd International Symposium on*, Aug 2015.
12. M. Sathiyanarayanan and D. Pirozzi, "Spherule diagrams: A matrix-based set visualization compared with euler diagrams," in *Green Computing and Internet of Things (ICGCIoT), 2015 International Conference on*, Oct 2015.
13. M. Sathiyanarayanan and D. Pirozzi, "Spherule diagrams with graph for social network visualization," in *Communication Systems and Networks (COMSNETS), 2016 8th International Conference on*, Jan 2016.

A Framework for Goal Compliance of Business Process Model

Dipankar Deb and Nabendu Chaki

Abstract In this paper, we propose a framework toward formal representation and validation of goal compliance for a business process model. All the tasks, postconditions, constraints, and goals are captured using first-order logic (FOL). We have used theorem prover (Prover9) for goal entailment. An experimental validation for goal compliance is presented considering a use case on health care domain. We start with an exhaustive solution space of all possible business process models for all possible activities on a particular domain and derive a reduced solution space of goal complied process models.

Keywords Business process modeling · Goal compliance · First-order logic

1 Introduction

In service sector, with increasing dynamics in business houses, there is always a steady demand for business process redesign. This in turn requires compliance of the redesigned business process depending on multiple criteria. This paper aims to achieve goal-based compliance for business process model.

In one of the recent works [4], an approach has been followed for searching the optimized business process model from the exhaustive space. The work in [4], however, does not mention the detail methodology of goal compliance from the exhaustive space. In this paper, we provide a framework that capture tasks, postcondition, constraints, and goal formally to check for goal entailment. We adopt a formal method using a use case as tasks, postcondition, constraints, and goal are not always quantifiable. The specific business process model which do not entails the

D. Deb (✉) · N. Chaki
Department of Computer Science and Engineering, University of Calcutta, Kolkata, India
e-mail: dipankar.deb@gmail.com

N. Chaki
e-mail: nabendu@ieee.org

P.K. Sa et al. (eds.), *Progress in Intelligent Computing Techniques: Theory, Practice, and Applications*, Advances in Intelligent Systems and Computing 719, DOI 10.1007/978-981-10-3376-6_38

goal will be pruned out of the exhaustive space, and thus providing with only goal satisfied process models in the exhaustive space.

Technically, we define our problem as: given a formal description of capability library T_i where each capability/task is annotated with postconditions E_i, set of constraints C_i, goal G, and a business process can be represented as a 4-valued tuple $<T_i, G, E_i, C_i>$, while a redesigned business process is represented as a 4-valued tuple $<T_j, G, E_j, C_j>$. Methodologically, we need to provide a validation of $<T_j, G, E_j, C_j>$ in terms of goal.

We have, $E_{cuf} = f(T_j, E_j, C_j)$, where E_{cuf} is the final cumulative effect for the sequence of T_j.

We need to determine $f(T_j, E_j, C_j) \vdash G$.

We consider the standards for the Indian Public Health system in the use case represented in Sect. 4 of this paper. Subsequently, we get the goal, postcondition, constraints, which are needed to comply for an optimized business process model in respect of the specific case like operation theater management. We map these into Prover9 using first-order logic (FOL) and validate whether specific business process model entails with the goal.

2 Related Work

The issue of service redesign is very vital because of its changing nature [6]. The issue of redesign is very much linked with business process compliance whose one of the criteria is business goal. So it is very important to check for goal satisfaction during the redesign of the service.

Most of the literature on business process redesign [11, 14], do not address the methodology of goal compliance for the improved process. A business process modeling framework proposed in [1] made it easy for IT people to understand and implement. The work in [15] presents a methodology toward modeling and validating an e-commerce system with a third-party payment platform. In another recent work presenting [16], a technique is proposed for modeling composite activities by including components of data, human actors, and atomic activities. The methodology supports representation of composite activities. A software tool is introduced in [12] for the automatic visualization of presents a software tool to automate visualization of the UML activity diagram. Modeling of medical services based on business process model is been described in [13]. A new modular workflow modeling language is proposed in [3] allows the designer to easily express data dependencies and time constraints. The literature survey above indicates the challenge of auditing the business process during redesign is need attention.

Verification of Business Process Constraints is been demonstrated in [5]. Various frameworks [6, 7] are developed to manage and check the violation of compliance policies by a given business process at design time, in order to minimize the cost of noncompliance. In [2] a semiautomated approach is discussed to synthesize business process templates out of compliance requirements. However, functional requirements are not considered as goal in [2].

The above survey reveals that very little work has been done toward developing a framework for efficient business process redesign that ensures goal compliance and auditing of business process using practical goal specification of functional requirements. This motivates us to work on generating business process design based on specific business logic incorporating the activities, postconditions, goals and constraints.

3 Proposed Methodology

The methodology for deducing a reduced exhaustive space with respect to work [4] is as follows:

1. Identify the business goals. Initiate goal reduction.
2. Identify tasks, postcondition, and constraints.
3. Identify roles and their responsibilities.
4. Formal analysis and interpretation of business goals, tasks, postconditions, and constraints.
5. Provide mechanism of cumulative accumulation of effect for each task and checking goal entailment of the final effect. The verification is done formally with theorem prover Prover9.

3.1 Tasks, Postcondition, and Constraints

In a business model, an actor is an entity which may be people or group of people. The actor is responsible for a role (combination responsibilities and action) guided by certain constraints.

Consider a case of "Pre-operation orders by surgeon for patient ready for operation" a hospital. Let e_1, e_2, e_3 and e_4 be the effect annotation at task t_1, t_2, t_3, and t_4, respectively. Let e_{cu1}, e_{cu2}, e_{cu3}, and e_{cu4} be the cumulative effect annotation at task t_1, t_2, t_3, and t_4 respectively. KB be the knowledge base which is nothing but a rule set (constraint).

$e_1 =$ *request for health report.*

$e_2 =$ *health report delivered to Surgeon.*

$e_3 =$ *Checking of health report by Surgeon.*

$e_4 =$ *Pre operation order for patient by surgeon.*

KB = Pre operation order are given after checking of health report by Surgeon.

We express the above informal representation formally as follows
$e_1 = request - report(x).$
$e_2 = deliver - report(x).$
$e_3 = check - report(x).$
$e_4 = preoperation - order(x).$
$e_{cu3} = request - report(x) \wedge deliver - report(x) \wedge check - report(x).$
$KB = (request - report(x) \wedge deliver - report(x) \wedge check - report(x))$
$\rightarrow preoperation - order(x).$

3.2 Business Goals

A business goal is a desired state of sequence of tasks. Goal can be described formally or informally. There is a need for formal representation of the goal for validating the process model. Sometimes objective and goals are represented informally for further reduction to subgoals formally. If G be the goal which can further business process, which can be further decomposed into subgoals $G_1, G_2, G_3 \ldots G_n$ such that $G = (G_1 \wedge G_2 \wedge G_3 \wedge \cdots \wedge G_n)$. Let us consider the case of operation theater and CSSD Management the goal is to “Make patient ready for OT” which can be AND-reduced to subgoals “Assign Surgeon to patient”, “Making report of patient available”, “Making pre-operation order”, “Get the patient admitted to hospital”, etc.

3.3 Goal Entailment Using Effect Accumulation

We accumulate immediate effect of each of the task to calculate the final effect of particular process model. Then we compare the final effect with goal. This can be done with goal entailment. Mathematically, let us consider F and G be the set of final effect scenarios after effects have been accumulated across all of the steps in a business process and the formal representation of the goals associated with the business process respectively. We require that the following constraint be satisfied: $F \vdash G$. If we again take the case of OT Management, we have seven tasks (do be done by surgeon) for making a patient ready for O.T. The work in [4], provides a methodology that produces the entire possible process model out of these seven tasks. Let us take a process model as shown in Fig. 1 for better understanding the methodology.

The corresponding pathways from Fig. 1 is as follows $<S, T_1, G_1, <T_2, G_2 < T_4, T_5 > T_3>, T_6, G_3, T_7>$.

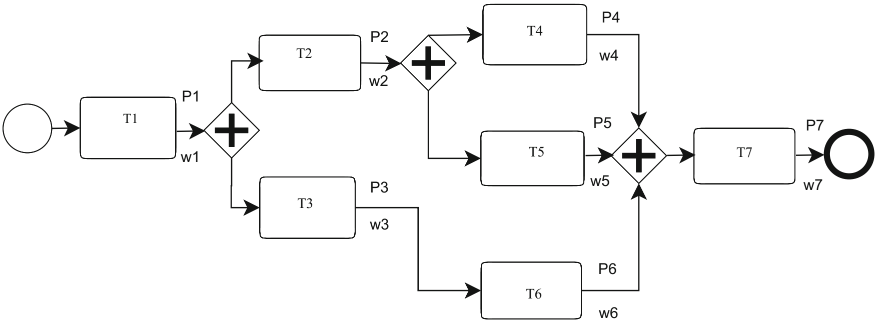


Fig. 1 A Process model derived where no of tasks = 7

These sequences in turn can be represented in CNL (Control Natural Language) and are exposed to the constraint satisfaction. The constraints defined by the analyst can be represented in CNL from the user friendly interface and will act as a knowledge base (KB).

There are three alternative effect scenarios during the cumulative effect at T_7. We proceed this way to gather final effect annotation at T_7. We write all cumulative effects and feed as assumption in Prover9 and check for goal entailment.

3.4 Proof for the Methodology

Definition 1: A scenario label [8] is a precise list of tasks that define a path leading from the Start Event in a model to the selected task. The simplest form of scenario label is a sequence of tasks.

Definition 2: Effect accumulation [8] involves the processing of immediate effect annotations for each of the tasks listed in the scenario label using a pair-wise operation where the immediate effect of S is combined with the immediate effect of T_1, the result being the cumulative effect at T_1. The cumulative effect at T_1 is then combined with the immediate effect of T_2 resulting in the cumulative effect at T_2, and so on up to T_n.

Definition 3: Role [10] involves a set of responsibilities and actions carried out by an actor or a group of actors within an organization.

Theorem: The final accumulated effect F entails a goal set G, i.e., $F \vdash G$, for every valid business process that meets the goal and is represented as a 4-tuple $<T, G, E, C>$ where T is the set of tasks, E is the set of effects, and C is the set of constraints.

Proof: Suppose G be the goal that can further business process which can be further decomposed into subgoals $G_1, G_2, G_3 \dots G_n$ such that $G = (G_1 \wedge G_2 \wedge G_3 \wedge \cdots \wedge G_n)$. $T_1, T_2, T_3 \dots T_n$ be the tasks and $e_1, e_2, e_3 \dots e_n$ be the effects annotated with the tasks $T_1, T_2, T_3 \dots T_n$ respectively. Let $c_1, c_2, c_3 \dots c_n$ be the constraints.

Let R be set of responsibilities and actions on tasks or activities which we call role. We define a relational operator ®between the roles and tasks.
Thus we have effect
$e_1 = R®T_1$
$e_2 = R®T_2$
Cumulative effect at task T_2 which immediately preceding T_1
$e_{2cu} = (R®T_1) \otimes (R®T_2)$
We can go on calculating to get the final accumulated effect $(R®T_1) \otimes (R®T_2) \otimes \cdots \otimes (R®T_n)$
There accumulation technique for the join or the split takes slightly different methodology but surely it the combination of Role and tasks.
We say that the process model whose cumulative effects matches the goals are the intended process models. Suppose for a valid process (we mean which meets the goal) we assume
$F| - G$ is false $\Rightarrow (e_1 \otimes e_2 \otimes e_3 \otimes \cdots \otimes e_n)| - G$ is false.
$\Rightarrow$ Either $e_1 \otimes e_2 \otimes e_3 \otimes \cdots \otimes e_n$ is not the final accumulated effect or G is not a goal.
Let us consider first $e_1 \otimes e_2 \otimes e_3 \otimes \cdots \otimes e_n$ is not the final accumulated effect
$\Rightarrow (R®T_1) \otimes (R®T_2) \otimes \cdots \otimes (R®T_n)$ is not true
$\Rightarrow$ Either Role or Tasks are not true, which is a contradiction.
Hence, we say that for a valid business process (we mean which meets the goal) F, i.e., the final accumulated effect entails the goal.

A smart interface can be provided for the analyst to make entry for goals, task, postcondition, and constraints. The entry is also restricted to specific format to the user, so that the formal specification of the entry can be recorded and is maintained in the database. We explicitly define the goal using task and constraints. The specifications that satisfy the goal (proof given by Prover9) are for the intended process and the rest are discarded from the exhaustive space.

4 Case Study

This section demonstrates the proposed methodology for the case of operation theater and CSSD management using the following steps:

Step 1: Informal representation of the task, postcondition and goal with respect to surgeon, staff nurse, and anesthetist for OT management.

Step 2: Formal representation written in first-order logic (FOL) of the task, postcondition and goal for OT management.

Step 3: Input the task, postcondition, and constraint in the assumption part and goal part in Prover9.

Step 4: Search for automated proof by Prover9.

Step 5: If proof is provided by Prover9, the business process model is goal complied, otherwise not.

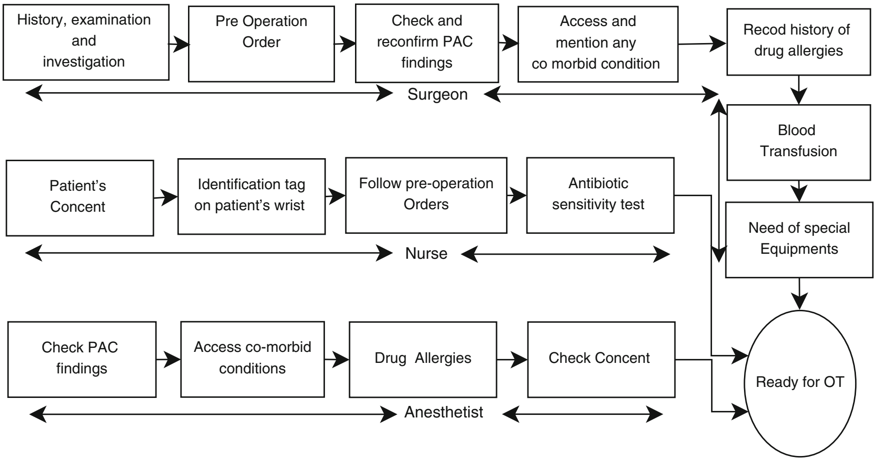

History, examination and investigation
Pre Operation Order
Check and reconfirm PAC findings
Access and mention any co morbid condition
Recod history of drug allergies
Surgeon
Blood Transfusion
Patient's Concent
Identification tag on patient's wrist
Follow pre-operation Orders
Antibiotic sensitivity test
Need of special Equipments
Nurse
Check PAC findings
Access co-morbid conditions
Drug Allergies
Check Concent
Ready for OT
Anesthetist

Fig. 2 Control flow diagram for making a surgical patient ready for OT

We identify the standard policies [9] of operation theatermor and CSSD management and for illustration purpose, we have identified policies for making a surgical patient ready for OT. A systematic flow of realization of making a surgical patient ready for OT is depicted in Fig. 2. The informal representation of the task, postcondition, and goal with respect to surgeon, staff nurse, and anesthetist for OT management are shown in Table 1. The corresponding formal representation written in first-order logic (FOL) of the task, postcondition, and goal with respect to surgeon for OT management are shown in Table 2. The representation with respect to staff nurse and anesthetists are omitted for brevity.

5 Implementation

We have mapped the goal, task, and constraint as input file to Prover9. If the cumulative effect of the tasks and constraint that is feed as assumption in Prover9 matches the goal, then it provides with a proof of goal entailment. If it fails to prove that indicates the cumulative effects of the tasks and constraint do not satisfies the goal. The screenshot of the proof for specific set activities, postcondition, and constraints is omitted due to space limit. The proof indicates that the business process is goal complied.

Table 1 Informal representation of the task, postcondition, and goal

Objective: make surgical patients ready for OT				
Repository		Roles		
		Surgeon	Nurse	Anesthetists
1	Task	History, examination and investigations	Patient consent to be taken	Check PAC findings
	Postcondition	History, examination and investigations are in hands	Signature	PAC findings reports are in hand
	Constraints	Required values of the examination and investigation meet certain threshold values	Patient/relative age >16	PAC findings below threshold level
2	Task	Preoperation orders	Identification tag on patient wrist	Assess co-morbid conditions
	Postcondition	Proper digestion/health condition	Patient has identification code	Reports are in hand
	Constraints	Patient is admitted in hospital	Patient is a registered in the hospital	Co-morbid conditions is negligible
3	Task	Check and reconfirm PAC findings	Follow preop orders	Check consent
	Postcondition	PAC findings reports are in hand	Proper digestion/health condition	Signature of the patient
	Constraints	PAC findings below threshold level	Patient is admitted in hospital	Patient/relative age >16
4	Task	Assess and mention any co-morbid condition	Antibiotic sensitivity test done	
	Postcondition	Reports are in hand	Report of Antibiotic sensitivity	
	Constraints	Co-morbid conditions is negligible	Antibiotic sensitivity should match with specified value	
5	Task	Record history of drug allergies		
	Postcondition	Records in hands of surgeon		
	Constraints	Availability of Laboratory database		
6	Task	Blood transfusion		
	Postcondition	Availability of Blood		
	Constraints	Blood bank data base available		
7	Task	Sister in charge of O. T. to be informed regarding the need for special equipment		
	Postcondition	Special equipment ready		
	Constraints	Sister in charge has information		

Table 2 Formal representation of the task, postcondition, and goal

Goal:		
∃ s,p,po,ad,r,pac,com,hisalg,data,blood,blooddata,e,i,c,s (assigned(s,p) ∧ has(p,r) ∧ order(p,po) ∧ get(p,ad) ∧ finding(p,pac) ∧ comorbidreport(p,com) ∧ hasdata(p,data) ∧ hashistory(p,hisalg) ∧ hasbloodavailibility(p,blood) ∧ hasbloodbank(p,blooddata) ∧ assignednurse(s,n) ∧ readyequipment(n,e) ∧ hasinformation(n,i) → ready(s,p))		
	Role	
	Surgeon	
1	Task	∃ s surgeon(s)
	Postcondition	∃ p patient(p)
	Constraint	∃ r report (r) ∃ s surgeon(s) → ∀ p (patient(p) → assigned(s,p)) ∀ p patient(p) → ∃ r (report(r) → has(p,r)) ∀ r report (r) ∃ p (patient(p) ∧ ¬ report(r) → ¬ ready(s,p)) ∀ r report (r) ∃ p (patient(p) ∧ report(r) → ready(s,p))
2	Task	∃ po preoporder(po)
	Postcondition	∃ ad hospitaladmission(ad)
	Constraint	∃ con patienthealth(con) ∀ p patient(p) → ∃ po (preoporder(po) → order(p,po)) ∀ p patient(p) → ∃ (ad hospitaladmission(ad) → get(p,ad)) ∀ ad hospitaladmission(ad) ∃ p (patient(p) ∧ ¬ hospitaladmission(ad)) → ¬ ready(s,p) ∀ ad hospitaladmission(ad) ∃ p (patient(p) ∧ hospitaladmission(ad)) → ready(s,p) ∀ con patienthealth(con) ∃ p (patient(p) ∧ ¬ con patienthealth(con)) → ¬ ready(s,p) ∀ con patienthealth(con) ∃ p (patient(p) ∧ con patienthealth(con)) → ready(s,p)
3	Task	∃ pac preoporder(pac)
	Postcondition	∀ p patient(p) → ∃ (pac preoporder(pac) → finding(p,pac))
	Constraint	∀ pac preoporder(pac) ∃ p (patient(p) ∧ ¬pac preoporder(pac)) → ¬ready(s,p) ∀ pac preoporder(pac) ∃ p (patient(p) ∧ pac preoporder(pac)) → ready(s,p)

(continued)

Table 2 (continued)

4	Task	∃ com comorbidcon(com)
	Postcondition	∀ p patient(p) → ∃ (com comorbidcon(com) → comorbidreport(p,com))
	Constraint	∀ com comorbidcon(com) ∃ p (patient(p) ∧ ¬com comorbidcon(com)) → ¬ready(s,p) ∀ com comorbidcon(com) ∃ p (patient(p) ∧ com comorbidcon(com)) → ready(s,p)
5	Task	∃ hisalg historyallergies(hisalg)
	Postcondition	∃ data database(data)
	Constraint	∀ p patient(p) → ∃ (data database(data) → hasdata(p,data)) ∀ p patient(p) → ∃ (hisalg historyallergies(hisalg) → hashistory(p,hisalg)) ∀ data database(data) ∃ p (patient(p) ∧ ¬data database(data) → ¬ready(s,p)) ∀ data database(data) ∃ p (patient(p) ∧ data database(data)) → ready(s,p) ∀ hisalg historyallergies(hisalg) ∃ p (patient(p) ∧ ¬hisalg historyallergies(hisalg)) → ¬ready(s,p) ∀ hisalg historyallergies(hisalg) ∃ p (patient(p) ∧ hisalg historyallergies(hisalg)) → ready(s,p)
6	Task	∃ blood bloodavailibility(blood)
	Postcondition	∃ blooddata bloodbank(blooddata)
	Constraint	∀ p patient(p) → ∃ (blood bloodavailibility(blood) → hasbloodavailibility(p,blood)) ∀ p patient(p) → ∃ (blooddata bloodbank(blooddata) → hasbloodbank(p,blooddata)) ∀ blood bloodavailibility(blood) ∃ p (patient(p) ∧ ¬blood bloodavailibility(blood)) → ¬ready(s,p) ∀ blood bloodavailibility(blood) ∃ p (patient(p) ∧ blood bloodavailibility(blood)) → ready(s,p) ∀ blooddata bloodbank(blooddata) ∃ p (patient(p) ∧ ¬blooddata bloodbank(blooddata) → ¬ready(s,p)) ∀ blooddata bloodbank(blooddata) ∃ p (patient(p) ∧ blooddata bloodbank(blooddata) → ready(s,p))

(continued)

Table 2 (continued)

7	Task	∃ n nurse(n)
	Postcondition	∃ e equipment (e)
	Constraint	∃ i information (i) ∀ s surgeon(s) → ∃ (nurse(n) → assignednurse(s,n)) ∀ p patient(p) → ∃ (equipment (e) → readyequipment(n,e)) ∀ n nurse(n) → ∃ (i information (i) → hasinformation(n,i)) ∀ n nurse(n) ∃ s surgeon(s) ∧ n nurse(n) → ready(s,p) ∀ n nurse(n) ∃ s surgeon(s) ∧ ¬ n nurse(n) → ¬ready(s,p) ∀ i information (i) ∃ n nurse(n) ∧ i information (i) → ready(s,p) ∀ i information (i) ∃ n nurse(n) ∧ ¬i information (i) → ¬ready(s,p)

6 Conclusion

In this paper, we have proposed a goal-based compliance framework for business process model redesign and supports modifications as per the goal. A healthcare-based use case has been considered to illustrate the methodology proposed. The validity for the proposed method has been proved formally as well as using a theorem prover tool called Prover9. The proposed framework can be extended for other compliance criteria for business process model, which remains as our future work.

References

1. Alotaibi, Y., Liu, F.: Business process modelling towards derive and implement it goals. In: Industrial Electronics and Applications (ICIEA), pp. 1739–1744. IEEE (2013)
2. Awad, A., Goré, R., Thomson, J., Weidlich, M.: An Iterative Approach for Business Process Template Synthesis from Compliance Rules. In: 23rd International Conference Advanced Information Systems Engineering, CAiSE 2011, pp. 406–421. Springer (2011)
3. Combi, C., Gambini, M., Migliorini, S., Posenato, R.: Representing business processes through a temporal data-centric workflow modeling language: An application to the management of clinical pathways. IEEE Transactions on Systems, Man, and Cybernetics: Systems 44(9), 1182–1203 (2014)
4. Deb, D., Chaki, N., Ghose, A.: Business process generation by leveraging complete search over a space of activities and process goals. In: Proceedings of the 5th International Conference on Cloud Computing and Services Science (CLOSER 2015), pp. 233–240. ScitePress (2015)
5. Gao, J., Chen, W., Wang, Y., Zhao, D., Li, W., Bo, Z.: Verification of business process constraints based on xyz/z. In: International Conference on Information Technology and Applications (ITA), pp. 479–482. IEEE (2013)
6. Ghose, A., Koliadis, G.: Auditing Business Process Compliance. In: Fifth International Conference Service-Oriented Computing - ICSOC 2007, pp. 169–180. Springer (2007)

7. Governatori, G., Milosevic, Z., Sadiq, S.: Compliance checking between business processes and business contracts. In: 2006 10th IEEE International Enterprise Distributed Object Computing Conference (EDOC'06), pp. 221–232 (2006)
8. Hinge, K., Ghose, A., Koliadis, G.: Process seer: A tool for semantic effect annotation of business process models. In: Enterprise Distributed Object Computing Conference, 2009. EDOC'09. IEEE International, pp. 54–63. IEEE (2009)
9. Indian public health standards, http://nrhm.gov.in (2012)
10. Koubarakis, M., Plexousakis, D.: A formal framework for business process modelling and design. Information Systems 27(5), 299–319 (2002)
11. Limam Mansar, S., Reijers, H.A., Ounnar, F.: Development of a decision-making strategy to improve the efficiency of bpr. Expert Systems with Applications 36(2),3248–3262 (2009)
12. Malesevic, A., Brdjanin D., Maric, S.: Tool for automatic layout of business process model represented by uml activity diagram. In: IEEE EUROCON, pp. 537–542. IEEE (2013)
13. Natalia, C., Alexandru M., Mihai, S., Stefan, S., Munteanu, C.: Medical services modelling based on business process model framework. In: IEEE E-Health and Bioengineering Conference (EHB), pp. 1–4. IEEE (2013)
14. Reijers, H.A., Liman Mansar, S.: Best practices in business process redesign: an overview and qualitative evaluation of successful redesign heuristics. Omega 33(4), 283–306 (2005)
15. Yu, W., Yan, C., Ding, Z., Jiang, C., Zhou, M.: Modeling and validating e-commerce business process based on petri nets. IEEE Transactions on Systems, Man and Cybernetics: Systems 44(3), 327–341 (2014)
16. Zhang, Y., Perry, D.: A goal-directed modeling technique towards business process. In: IEEE 8th International Symposium on Service Oriented System Engineering (SOSE), pp. 110–121. IEEE (2014)

Comparative Analysis of Adaptive Beamforming Techniques

Smita Banerjee and Ved Vyas Dwivedi

Abstract Adaptive beamforming (ABF) techniques are used to produce higher gain in the user directions and lower gain in the interferer directions by calculating the excitation weights. It tries to reduce the difference between the desired and actual signal and maximize the signal-to-interference ratio (SIR). But in severe interference environment when the actual signal is weak, the effect of SIR on the radiation pattern needs to be considered. This paper describes the effect of signal-to-interference ratio on different adaptive beamforming techniques such as non-blind least mean square (LMS) and evolutionary particle swarm optimization (PSO). The performance and validation of beamforming algorithms are studied through MATLAB simulation by varying SIR parameter for desired and interference direction. Different weights are obtained using this beamforming algorithm to optimize the radiation pattern. The parameters for comparison are the main beam and null placement keeping signal-to-noise (SNR) constant for specified user and interferer.

Keywords Adaptive antenna · Adaptive beamforming · Particle swarm optimization · Least mean square · Signal-to-interference ratio · Excitation weights

1 Introduction

In satellite communication systems, the receiver receives extremely weak signals from the satellite. Hence, adaptive array signal processing technology is used to improve received radiation patterns quality according to the changing environment. An adaptive antenna is an antenna array with amplitude and phase feedback control to change its received radiation pattern so as to point the reception in a certain

S. Banerjee (✉)
School of Engineering, RK University, Rajkot, India
e-mail: smita161@gmail.com

V. Vyas Dwivedi
C.U. Shah University, Wadwan City, India
e-mail: vedvyasdwivedphd@gmail.com

P.K. Sa et al. (eds.), *Progress in Intelligent Computing Techniques: Theory, Practice, and Applications*, Advances in Intelligent Systems and Computing 719, DOI 10.1007/978-981-10-3376-6_39

direction and discards from other direction. The need to eliminate, remove, or reduce the undesired signal effect as compared to the desired one motivates more research in adaptive beamforming techniques [1–11].

There are different adaptive beamforming algorithms studied in literature which are used in the adaptive antenna array [12–24]. Beamformers are analyzed and compared based upon statistically optimum blind and non-blind adaptive beamforming for beamforming capability and convergence rate. It is observed that the convergence rate of least mean square (LMS) is slowest where as constant CGM is the fastest among all. SMI is found to have more computational complexity. Recursive least square (RLS) is found to have higher side lobe level (SLL) and null depths as compared to CGM [16]. It was observed that the conventional adaptive beamforming (ABF) technique like minimum variance distortionless response (MVDR) improves the signal-to-interference-plus-noise ratio (SINR) but unable to reduce the SLL [17]. Hence to improve the SINR with reduced SLL, many optimization techniques have been used in ABF application. Adaptive Mutated Boolean Particle Swarm Optimization (AMBPSO) technique takes the uncorrelated desired and interferer signal directions and succeed in providing good SINR value with lower SLL as compared to conventional MVDR [18]. Adaptive Dispersion Invasive Weed Optimization (ADIWO) shows faster and better SLL as compared to PSO and improvement in the capability to move the major lobe and the null point [19]. Hybrid particle swarm optimization with Gravitational Search Algorithm (Hybrid PSOGSA) shows its ability for optimization in beamforming for a larger number of user signals and speedy computation using parallel GSA as compared to sequential stand alone algorithms but cannot maximize the gain along the user direction [20, 21]. Mementic algorithm shows optimal radiation pattern design to maximize the signal-to-interference ratio (SIR) by perturbing the phase position [22].

In all of the above adaptive beamforming techniques proposed so far try to reduce the difference between the desired and actual signal and maximize the signal-to-interference ratio (SIR). But in severe interference environment when the actual signal is weak, the effect of SIR on the radiation pattern needs to be considered. The present study analyzes different adaptive techniques such as non-blind LMS and evolutionary PSO through MATLAB simulation by varying SIR parameter for desired and interference direction. Different weights are obtained using this beamforming algorithm to optimize the radiation pattern. The parameters for comparison are the main beam and null placement keeping signal-to-noise (SNR) constant. The mean SLL and directivity are also studied.

2 Adaptive Beamforming Problem Formulations

An ULA will receive the incident signals which are multiplied by the amplitude and phase weight of antenna elements. These are summed to produce the array output in the form of received signal. The received far field radiation pattern of the linear array is represented in terms of array factor (AF) by [14],

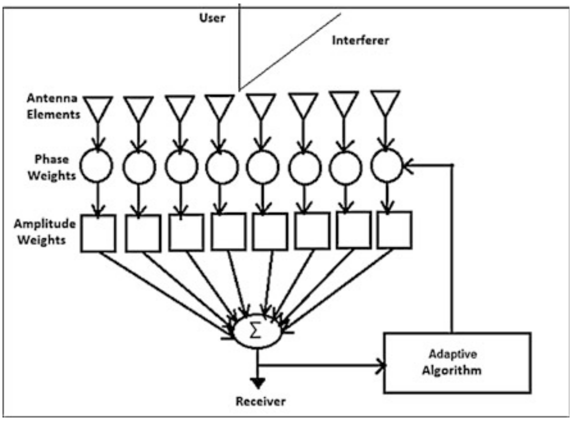


Fig. 1 Block diagram of adaptive antenna array

$$AF = \sum_{n=1}^{N} X(k) \ {*}w_n, \tag{1}$$

where N is the number of antenna elements, $w_n = a_n * \exp(jb_n)$ = complex array weights at element n, a_n is the amplitude weight at element n, b_n is the phase shift weight at element n.

The radiation pattern of ULA is controlled through various adaptive algorithms. It will compare the output received signal radiation pattern with the desired radiation pattern. If the received signal is not as the desired one, then adaptive algorithm will alter the weights in order to reduce the error until both the pattern remains same. The received array pattern is given in the feedback loop to optimize the radiation pattern. The objective is to provide more gain in the desired signal direction and lesser gain in the interferers' direction. Figure 1 shows the block diagram of an adaptive antenna array.

3 Adaptive Beamforming Using Particle Swarm Optimization

Particle swarm optimization (PSO) was developed by Eberhart and Arora [25, 26]. It is used as adaptive algorithm to search the optimized adaptive antenna radiation pattern. This is done using the algorithm summarized in the Table 1. The amplitudes excitations are kept constant whereas the phase excitations are selected as the optimization parameters. Hence the AF can be written as

$$AF = \sum_{n=1}^{N} X(k) \ * \ \exp^{jb_n} \tag{2}$$

Table 1 Algorithm for adaptive beamforming using PSO

Step 1: Initialize population, number of iterations (ni), tuning parameters ($\varphi 1$ and $\varphi 2$) and weights (w). The particle corresponds to phase bn in the interval [−2π, 2π].
Step 2: Initialize the value of the kth variable in the population by $b_n(ni,k) = b_n(ni,\min) + (b_n(ni,\max) - b_n(ni,\min))u(ni)$, where k = 1, 2, …, npop and u(ni) is the arbitrary number chosen between 0 and 1. Initialize the velocities of the kth variable as v (ni, k) = 0.
Step 3: Estimate the fitness function (FF) for each particle bn(ni). Compute FF (ni, k) as per the Eq. (7).
Step 4: Calculate personalbest (ni, k) = FF(ni, k) and globalbest (ni) = max (personalbest (ni, k)) with its location personalbest (k) and globalbest.
Step 5: Modify velocity v (ni + 1, k) and position bn (ni + 1, k) using $v(ni+1,k) = w *v(ni,k) + \phi 1(p(b_n ik) - b_n(ni,k))u(ni) + \varphi 2(g(ib_n)$ $- b_n(ni,k))u(ni)$ $b_n(ni+1,k) = b_n(ni,k) + v(ni+1,k)$
Step 6: Modify fitness function FF (ni + 1, k).
Step 7: If FF (ni + 1, k) > FF (ni, k), then personalbest (ni + 1, k) = FF(ni + 1, k).
Step 8: Modify globalbest (ni + 1, k) = max (personalbest (ni + 1, k)).
Step 9: If i < imax then ni + 1 and move to step (5), or else stop.

The objective function is formulated to find the values of phase of the element of antenna array to direct the major lobe towards the desired user while low gain towards interfering user. It is formulated using the AF equation. For 1 user and 2 interferer,

$$\text{Fitness function for Beamforming} = \mathrm{AF}(\theta_{s1}) - [\mathrm{AF}(\theta_{i1}) + \mathrm{AF}(\theta_{i2})], \tag{3}$$

where

$$AF(\theta_{s1}) = \sum_{n=1}^{N} \exp^{-j\pi(n-1)(\sin\theta_{s1})} * \exp^{jb_n}$$

$$AF(\theta_{i1}) = \sum_{n=1}^{N} \exp^{-j\pi(n-1)(\sin\theta_{i1})} * \exp^{jb_n} \text{ and } AF(\theta_{i2}) = \sum_{n=1}^{N} \exp^{-j\pi(n-1)(\sin\theta_{i2})} * \exp^{jb_n} \tag{4}$$

4 Adaptive Beamforming Using Least Mean Square Algorithm

Widrow and Hoff developed least mean square (LMS) algorithm in 1960. The optimum weights can be estimated with LMS algorithm. The algorithm recursively calculates and modifies the weight vector between the beamformer output and the desired signal as summarized in Table 2 [13].

5 Numerical Simulation Results

PSO and LMS were applied on a 16-element ULA with $\lambda/2$ interelement spacing and were compared on the basis of the SIR. The simulations are done using MATLAB. All the algorithms were executed for 200 iterations and the termination criterion is set for the number of iterations. For PSO, the population size is assumed

Table 2 Algorithm for adaptive beamforming using LMS

Step 1: Initialize number of iteration nimax and the value of μ.
Step 2: Initialize weight WLMS, error ELMS and output yLMS as 0.
Step 3: Calculate Output, yLMS (ni, k) = WLMS(ni, k)Hx(k)
Step 4: Calculate Error, ELMS (ni, k) = Su(k) − yLMS (ni, k)
Step 5: Compute Weight, WLMS (ni + 1, k) = WLMS (ni, k) + μx(k)ELMS*(ni, k)
Step 6: If ni > nimax, in that case stop, or else move to step (3) to update output, error and weight.

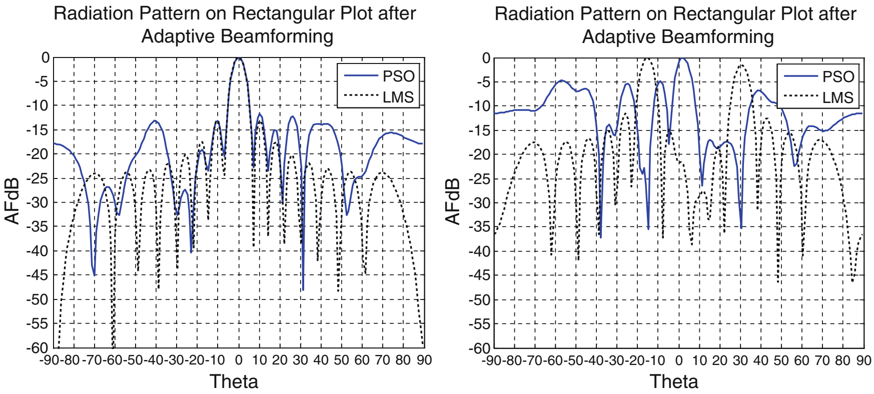


Fig. 2 Best radiation pattern found by PSO and LMS for 16 element antenna array with user at 0° and interferers at −15° & 30° with SNR = 30 dB. **a** Rectangular plot for SIR = 30 dB ($SLL_{PSO} = -15.15$ dB, $SLL_{LMS} = -19.12$ dB, Directivity = 6 dB). **b** Rectangular plot for SIR = −30 dB ($SLL_{PSO} = -10.35$ dB, Directivity = 3 dB)

Table 3 Optimized excitation weights for SIR = 30 dB

N	WPSO	WLMS	N	WPSO	WLMS
1	1.0000 + 0.0000i	1.0000 + 0.0000i	9	−0.2931 − 0.9561i	1.0006 − 0.0033i
2	0.8406 − 0.5416i	1.0015 + 0.0038i	10	−0.9264 − 0.3765i	0.9996 + 0.0059i
3	0.5921 + 0.8059i	1.0067 + 0.0007i	11	−0.1483 + 0.9889i	1.0025 − 0.0061i
4	0.9996 + 0.0271i	0.9957 − 0.0074i	12	−0.7944 + 0.6075i	0.9967 − 0.0080i
5	0.5373 − 0.8434i	0.9941 − 0.0078i	13	−0.1794 + 0.9838i	0.9940 − 0.0060i
6	−0.0491 − 0.9988i	1.0019 − 0.0054i	14	−0.6254 − 0.7803i	1.0029 − 0.0059i
7	0.6675 − 0.7446i	1.0065 − 0.0069i	15	0.8266 − 0.5627i	1.0095 − 0.0096i
8	0.5434 + 0.8395i	1.0024 − 0.0072i	16	0.2101 − 0.9777i	1.0056 − 0.0080i

as 100 and tuning parameter $\varphi 1$ and $\varphi 2$ are set to 2.0. Phase excitation bn is chosen as the design variable in the PSO with lower and upper limit taken in the range of [−2π, 2π] with initial values of position and velocities are taken as random. For LMS, μ is taken as 0.001 and the initial weight and error are set to 0.

Based upon the aims to maximize the AF gain of the desired user and minimize the AF gain of the interfering user. The ULA receives a desired signal arriving from angle $\theta_{s1} = 0$ and 2 interference signals arriving from angles $\theta_{i1} = -15$ and $\theta_{i2} = 30$. Two cases are studied for SIR = 30 dB and −30 dB values keeping SNR = 30 dB.

For each case, it was observed that PSO algorithm produce main lobe along θ_{s1} and nulls towards θ_{i1} and θ_{i2}. The AF gain along the main lobe is 0 dB whereas the AF gain towards the null is −22 to −48 dB as shown in Fig. 2a. LMS algorithm also produces main lobe gain of 0 dB along the θ_{s1} direction and null gain of −33 to −40 dB for SIR = 30 dB. As SIR reduces to −30 dB, LMS fail to point the main beam and null along the user and the interferer direction as shown in Fig. 2b. Tables 3 and 4 give the optimized excitation weights for PSO and LMS for SIR = 30 dB and SIR = −30 dB.

Table 4 Optimized excitation weights for SIR = −30 dB

N	WPSO	WLMS	N	WPSO	WLMS
1	1.0000 + 0.0000i	1.0000 + 0.0000i	9	−0.9984 + 0.0559i	0.8813 + 0.1116i
2	0.6505 + 0.7595i	−0.3709 + 0.2632i	10	−0.0746 + 0.9972i	−0.5332 + 0.2516i
3	−0.9819 + 0.1894i	−0.8997 + 0.9081i	11	−0.3553 + 0.9347i	−1.0053 + 0.7838i
4	0.0823 + 0.9966i	−0.2752 + 0.3665i	12	0.2483 − 0.9687i	−0.2550 + 0.2055i
5	0.5657 − 0.8246i	−0.0522 − 1.0152i	13	−0.5910 + 0.8067i	0.0815 − 1.1111i
6	−0.8744 − 0.4853i	−0.3545 − 1.1998i	14	0.0308 − 0.9995i	−0.1963 − 1.1714i
7	0.6679 + 0.7443i	0.1701 − 0.0844i	15	0.9996 − 0.0285i	0.2579 + 0.0495i
8	0.5761 + 0.8174i	1.1847 + 0.4617i	16	−0.2219 − 0.9751i	1.1464 + 0.6194i

6 Conclusions

In this paper, ABF based on PSO and LMS method has been simulated for 16 elements ULA. A performance analysis and validation is done by changing the values of SIR for specific user and interferer position. The main lobe gain and null depth are calculated to validity this approach. It is shown that the PSO-based beamformer provides accurate 0 dB main beam gain and null depth of −22 to −48 dB with better SLL for each case of SIR. However, LMS shows better SLL than PSO but fail to provide main beam and null placement reduced value of SIR. Therefore, the PSO method seems to be simple and appropriate in ABF applications based on the fitness function. ABF using PSO shows mean side lobe level (SLL) of −15 dB with a directivity of 6 dB for SIR = 30 dB. It can be further studied with complex fitness functions in order to improve the value of SLL.

References

1. Balanis. C.A.: Antenna Theory: Analysis and Design. 3rd Edition, John Willy & Sons Inc., New York (2005).
2. Das, S.: Smart antenna design for wireless communication using adaptive beam-forming approach, Proceedings of the IEEE Region 10 TENCON Conference, (2008) 19–21.
3. Lian, K. J.: Adaptive antenna arrays for satellite personal communication systems, Master of Science Thesis, Virginia polytechnic institute and state university, Blacksburg, (1997).
4. Canabal, A., Jedicka, R. P., Pino, A. G.: Multifunctional phased array antenna design for satellite tracking, Elsevier Journal, Vol. 57(12), (2005) 887–900.
5. Jiancheng, W., Hui, Q., Jianming, C.: Optimizing adaptive linear array antenna pattern under intensive interference environment using Genetic Algorithm, Proceedings of the 4th International Conference on Intelligent Computation Technology and Automation, (2011) 185–187.
6. Banerjee, S., Dwivedi, V. V.: Linear Antenna Array Synthesis to Reduce the Interference in the Side Lobe using Continuous Genetic Algorithm Proceedings of the IEEE 5th International Conference on Advances in Computing and Communications, (2015) 291–296.

7. Hossain, S., Islam, M. T., Serikawa, S.: Adaptive Beamforming Algorithms for Smart Antenna Systems, Proceedings of the International Conference on Control, Automation and Systems, Coex, Seoul, Korea, (2008).
8. LianBaowang, Z. H., Juan, F.: Adaptive Beamforming Algorithm for Interference Suppression in GNSS Receivers, International Journal of Computer Science & Information Technology, Vol. 3(5), (2011) 17–28.
9. Widrow, B.: Adaptive antenna systems, Proceeding of the IEEE, Vol. 55(12), (1967).
10. Applebaum, S. P.: Adaptive Arrays, IEEE Transactions on Antennas and Propagation, (1976).
11. Gu, Y. J., Shi, Z. G., Chen, K. S., Li, Y.: Robust Adaptive Beamforming for Steering Vector Uncertainties Based on Equivalent DOAs Method, Progress In Electromagnetics Research, Vol. 79, (2008) 277–290.
12. Kamboj, S., Dahiya, R.: Adaptive antenna array for Satellite Communication Systems, Proceedings of the International Multi Conference of Engineers and Computer Scientists, Hong Kong, (2008).
13. Banerjee, S., Dwivedi, V. V.: An LMS Adaptive Antenna Array", International Journal of Advanced Research in Engineering and Technology, Vol. 4(6), (2013) 166–174.
14. Banerjee, S., Dwivedi, V. V.: Review of adaptive linear antenna array pattern optimization, International Journal of Electronics and Communication Engineering, Vol. 2(1), (2013) 25–42.
15. Banerjee, S., Dwivedi, V. V.: Linear array synthesis using Schelkunoff polynomial method and particle swarm optimization, Proceedings of the IEEE International Conference on Advances in Computer Engineering and Applications, (2015) 727–730.
16. Saxena, P., Kothari, A. G.: Performance Analysis of Adaptive Beamforming Algorithms for Smart Antennas, Proceedings of the IERI International Conference on Future Information Engineering, Vol. 10, (2014), 131–137.
17. Liu, F., Wang, J., Sun, C. Y., Du, R.: Robust MVDR beamformer for nulling level control via multi-parametric quadratic programming, Progress In Electromagnetics Research C, Vol. 20, (2011) 239–254.
18. Zaharis, Z. D., Yioultsis, T. V.: A Novel Adaptive Beamforming Technique Applied on Linear Antenna Arrays using Adaptive Mutated Boolean PSO, Progress In Electromagnetics Research, Vol. 117, (2011), 165–179.
19. Zaharis, Z. D., Skeberis, C., Xenos, T. D.: Improved Antenna Array Adaptive Beam-Forming with Low Side Lobe Level using A Novel Adaptive Invasive Weed Optimization Method, Progress In Electromagnetics Research, Vol. 124, (2012) 137–150.
20. Magdy, A., EL-Ghandour, O.M., Hamed, H.F.A.: Performance Enhancement for Adaptive Beam-Forming Application Based Hybrid PSOGSA Algorithm. Journal of Electromagnetic Analysis and Applications, Vol. 7, (2015) 126–133.
21. Magdy, A., EL-Ghandour, O.M., Hamed, H.F.A.: Adaptive Beam-forming Optimization Based Hybrid PSOGSA Algorithm for Smart Antennas Systems, Progress In Electromagnetics Research Symposium Proceedings, Prague, Czech Republic, (2015) 973–977.
22. Hsu, C. H., Shyr, W. J.: Memetic Algorithms for Optimizing Adaptive Linear Array Patterns By Phase-Position Perturbations, Circuits Systems Signal Processing, Vol. 24(4), (2005) 327–341.
23. Zuniga, V., Erdogan, A. T., Arslan, T.: Adaptive radiation pattern optimization for antenna arrays by phase perturbations using particle swarm optimization, IEEE NASA/ESA Conference on Adaptive Hardware and Systems (AHS), (2010) 209–214.
24. Rao, A. P., Sarma, N. V. S. N.: Interference Suppression in Multiple Beams Adaptive Linear Array using Genetic Algorithm, Antenna Test and Measurement Society, India, (2011).
25. Eberhart, R. C., Lu, Y.: Particle swarm optimization: developments, applications and resources, evolutionary computation, Proceedings of the 2001 Congress on Evolutionary Computation, Vol. 1, (2001) 81–86.
26. Arora, R. K.: Optimization: Algorithms and Applications, 1st Edition, CRC Press, New York (2015).

Part VI
Internet, Web Technology, Web Security, and IoT

Securing an External Drive Using Internet with IOT Concept

Rajneesh Tanwar, K. Krishnakanth Gupta and Purushottam Sharma

Abstract In the computer world, data security is the main issue. Many attacks and techniques are used for stealing data. Instead of considering on internet, o ine storage like in hard disk, etc., should also get secured enough so that in case of theft, user information should not get accessed. Till now, third party software is available for securing but they all are easily cracked. For securing hard disk in such a way that attacker not get information from device, a small chip will be mounted inside the hard disk which is always concatenated with user email and always sends location and information of the attached machine. It also contains double password security and can also receive message from concatenated email for blocking or encrypting the data inside the disk. This approach is like using IOT for securing data in hard disk.

Keywords Hard disk · Third party software · Double password security · Encryption · Internet of things (IOT)

1 Introduction

In today's world, for storing huge amount of data any storage unit of very low cost and high storing size is available in the market. High volume products are nowadays in need of a common person or citizen as the data is rowing in very big ratio [1]. Despite the importance of the subject, there are many techniques or paper that are formatted and formulated for the protection of these devices or protecting data

R. Tanwar (✉) · K. Krishnakanth Gupta · P. Sharma
Department of Information Technology, Amity University,
Noida, UP, India
e-mail: rajneeshtanwar15@gmail.com

K. Krishnakanth Gupta
e-mail: krishnakanthgupta@outlook.com

P. Sharma
e-mail: psharma5@amity.edu

P.K. Sa et al. (eds.), *Progress in Intelligent Computing Techniques: Theory, Practice, and Applications*, Advances in Intelligent Systems and Computing 719, DOI 10.1007/978-981-10-3376-6_40

stored inside them [2]. Their data are typically based on extrapolation from accelerated life test data of small populations or from returned unit databases. Advancement in life has also made advancement in the technology but till now no good solution for protecting such devices is available in the market [3]. Disk drives are generally very reliable but they are also very complex components. This basically defines that these devices are very hard and not easy get fails in change of environment. These are very reliable due to which every citizen keeps confidential data in it but the main concern is security in which these device have very lose hand. In this paper, we are going to build up these device smart enough so that their lose hand in security will become strongest part by saving data to be accessed by unauthorized one [4].

2 Proposed Methodology

This method works whenever we connect drives to the computer, it installs drivers and at that time it will install one more program in parallel manner which will get the information of computer like IP address and system information connected to the computer by the user. This information is stored at cloud so administrator of that drive can see the information all the time. By the help of your system, it sends the information through the computer by assigning port number, we design assigning of port number so whenever the external drive connects to the computer then port must be awake or it should be in sleep mode so that no one can intrude the system and no security issues can arise. If the system is not connected to Internet then it should hold the data of the system which was collected and should be hidden in the computer so no one can see it. And the collected raw information should be created into knowledge using of big data so that administrator can have control over the drive by locking any specific folder or locking entire disk with more encryption techniques. So in this way we can control the drive over internet.

Here we should concentrate on one more thing which is how to get IP address and mac ID of a system, though generally we use ipconfig and ifconfig commands in windows and linux operating systems to get ip address of the system, and getmac is used to get mac ID of a system so we include this commands by creating batchfiles with in the driver software so it can easily get information, here we should be concerned about one thing that is executing batchfiles system may lead to malicious malware so, considering that we should create batch files which do not harm the system (Fig. 1).

Flow diagram shown in Fig. 2 describes the flow of working of hard disk in detection of the attached machine and also for encrypting and decrypting information in disk according to the need. In this, whole working of mounted CHIP and hark disk is described to know how intelligently all information of the machine will be received by user when attached.

In case if the hard disk is lost or stolen, the flow will change remain same and encryption of data will be done by the following steps and flow. Figure 3 describes

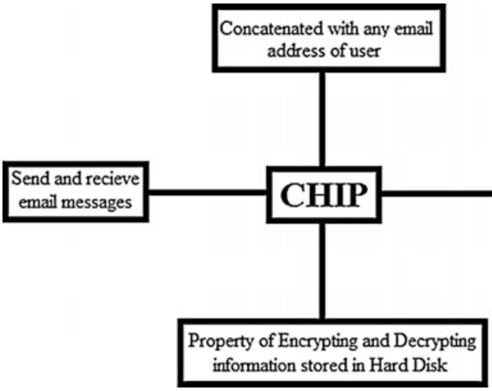

Concatenated with any email address of user
Send and recieve email messages
CHIP
Store information of attached macine and password for authentication
Property of Encrypting and Decrypting information stored in Hard Disk

Fig. 1 Property of mounted CHIP

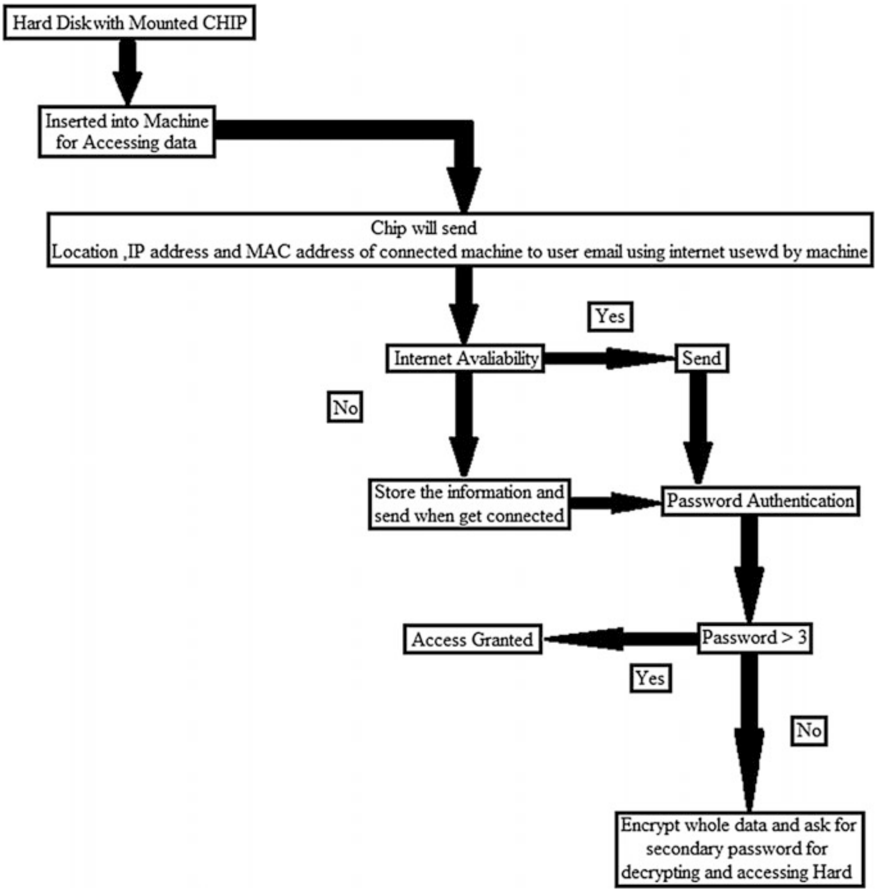

Hard Disk with Mounted CHIP
Inserted into Machine for Accessing data
Chip will send
Location ,IP address and MAC address of connected machine to user email using internet usewd by machine
Yes
Internet Avaliability
Send
No
Store the information and send when get connected
Password Authentication
Access Granted
Password > 3
Yes
No
Encrypt whole data and ask for secondary password for decrypting and accessing Hard

Fig. 2 Flow of working of CHIP with hard disk

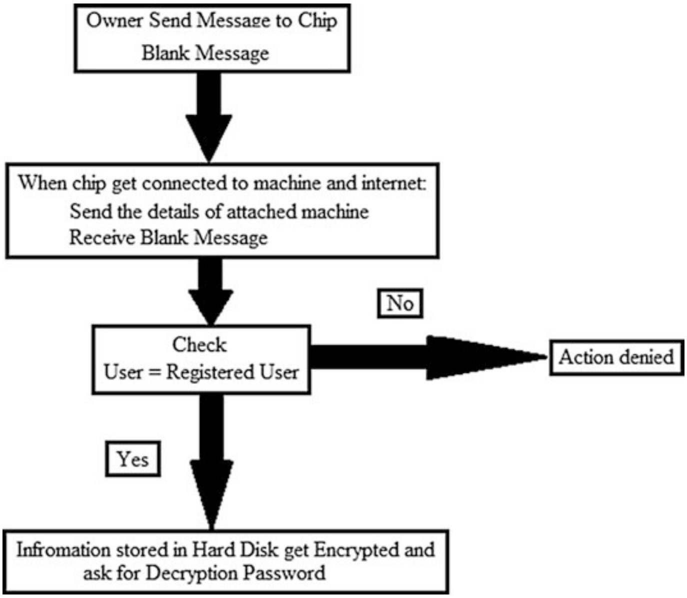

Owner Send Message to Chip
Blank Message
When chip get connected to machine and internet:
Send the details of attached machine
Receive Blank Message
No
Check
User = Registered User
Action denied
Yes
Infromation stored in Hard Disk get Encrypted and ask for Decryption Password

Fig. 3 In case of hard disk lost or theft

the flow or working of hard disk in case of lost disk or stolen disk. How disk will automatically encrypt the data inside disk so that unauthorized user will not able to get the disk information.

3 Conclusion

In this paper, we propose how to secure external hard drive through IOT method so that external drive can be protected even if it was robbed and giving control over the disk by use of computer if it is connected. By this way, whole data in the disk will be safe and all safety measures are taken by user only and no third party software are used for performing such security in hard disk. The proposed system is capable of catching the location of hard disk in case of theft and securing the internal data by encrypting inside hard disk when inserting wrong password.

References

1. Peter Lyman and Hal R.Varian. How much information? October 2013.
2. Dave Anderson, Jim Dykes, and Erik Riedel. More than an interface - scsi vs. ata. In Proceedings of the 2nd USENIX Conference on File and Storage Technologies (FAST03), pages 245 257, February 2003.

3. Jon G. Elerath and Sandeep Shah. Server class disk drives: How reliable are they? In Proceedings of the Annual Symposium on Reliability and Maintainability, pages 151 156, January 2004.
4. Eduardo Pinheiro, Wolf-Dietrich and Luiz Andre Barroso. Failure Trends in a Large Disk Drive population, proceeding of the 5th USENIX conference on the File and Storage Technology, February 2007.
5. Storage View, http://www.storagereview.com/ssd_vs_hdd.
6. Computer Weekly, http://www.computerweekly.com/feature/Self-encrypting-drives-SED-the-best-kept-secret-in-hard-drive-encryption-security.

An Intelligent Algorithm for Automatic Candidate Selection for Web Service Composition

Ashish Kedia, Ajith Pandel, Adarsh Mohata and S. Sowmya Kamath

Abstract Web services have become an important enabling paradigm for distributed computing. Some deterrents to the continued popularity of the web service technology currently are the nonavailability of large-scale, semantically enhanced service descriptions and limited use of semantics in service life cycle tasks like discovery, selection, and composition. In this paper, we outline an intelligent semantics-based web service discovery and selection technique that uses interfaces and text description of services to capture their functional semantics. We also propose a service composition mechanism that automatically performs candidate selection using the service functional semantics, when one web service does not suffice. These techniques can aid application designers in the process of service-based application development that uses multiple web services for its intended functionality. We present experimental and theoretical evaluation of the proposed method.

Keywords Web services composition · Semantic search · NLP

1 Introduction

W3C defines a web service as a software system that supports interoperable machine to machine interaction over a network. A web service can be uniquely identified by a URI and each is described using one or more XML-based documents which define

A. Kedia (✉) · A. Pandel · A. Mohata · S. Sowmya Kamath
Department of Information Technology, National Institute of Technology Karnataka,
Surathkal, Mangalore, India
e-mail: ashish1294@gmail.com

A. Pandel
e-mail: ajithpandel@gmail.com

A. Mohata
e-mail: amohta163@gmail.com

S. Sowmya Kamath
e-mail: sowmyakamath@nitk.ac.in

P.K. Sa et al. (eds.), *Progress in Intelligent Computing Techniques: Theory, Practice, and Applications*, Advances in Intelligent Systems
and Computing 719, DOI 10.1007/978-981-10-3376-6_41

the service interfaces, the functionality provided by it and also prescribe the manner in which it interacts with other systems. Large-scale applications can be easily built by composing loosely coupled web services [10], thus enabling a service oriented architecture.

While developing applications, searching for an appropriate web service that can provide a required functionality is not trivial. Web service discovery and retrieval often becomes a bottleneck. Recent years have not only seen an explosive increase in the number of web services being offered but have also witnessed a rise in number of standards to describe those service. The problem is further compounded by the fact that there is no central repository with all service descriptions. Web service standards such as UDDI (Universal Description, Discovery and Integration) which relied on a central registry of all web services are now obsolete owing to its low benefit/complexity ratio. The requirements have also escalated. Developers now need a method to dynamically look up for appropriate web service during run-time making service discovery a challenging task. Semantic web technology attempts to automate the web service discovery. Most of the existing algorithms for automated web service discovery serves to only web services that have explicit semantic tags associated with their description document which is an unreasonable expectation. A large number of existing web services do not have any semantics tags associated with their description document. Approaches to convert existing non-semantic description documents of web services to corresponding semantic ones are also severely limited.

Our work is focused on studying the existing methods of discovering web services and develop a method to automatically index a set of web services using their description documents such that services can be automatically searched and composed based on user's need. The rest of this paper in organized as follows. In Sect. 2 we will discuss the existing work concerning the described problem. This is followed by Sect. 3 that describes methodologies to index web services using their description documents. Section 4 talks about the algorithm to search the indexed web services to find services relevant to the user. In Sect. 5 we propose a methodology to automatically compose multiple web services. In Sect. 6 we will discuss the results obtained and analyze the proposed method. Finally, Sect. 7 concludes our work with a few possible future improvements.

2 Related Work

Yanbin et al. [7] have modeled the service discovery problem as an assignment problem using functional constraints. They have proposed an automatic semantic search algorithm which is loosely based in assignment algorithm. It uses 3 step match making - Service Library Matchmaking, Service Matchmaking, and Operation Matchmaking. Operation matchmaking can be further divided into interface matchmaking and concept matchmaking.

Platzer and Dustdar [8] proposed the construction of a vector space to index descriptions of already existing services. They have used the prevalent information retrieval methods over the existing standards to create a multidimensional “term space”, where each dimension represents a category of web services and then represent each web service in this space using a vector. The relative position of these vectors in the said space is used to compute the effective similarity of the corresponding web services.

Cuzzocrea et al. [1] have considered both internal structure and component of web services. They have outlined an algorithm for service discovery that represents composite OWL-S (Web Ontology Language for Services) documents using graphs. They proposed an algorithm that matches a group of services with a query using such graph-based representation. They have not only considered the similarity of individual services in the matched group but have also taken into account the flow or control between different services of that group.

Sangers et al. [9] have used popular NLP techniques like lemmatization, tagging parts of speech, and word sense disambiguation to establish the semantics of web service description. They also determine the senses of the relevant words in user’s query and then carry out a match-matching process between users query and indexed web services. In this method, a context aware search is performed, i.e., actual user’s need is matched with services that perform required computation. Fethallah et al. [2] have outlined a mechanism to use the external interface (inputs/outputs) of web services. They have used domain ontology to classify service interfaces and then index the corresponding service conceptually. Once the services are indexed they use the popular cosine similarity measurement to computer the degree of similarity between the query and the indexed services. The method yielded good result and is relative less resource intensive than the other existing methods.

Vector space search engine seems like a promising approach however it fails to account for the service semantics which is an important parameter in service discovery. Our algorithm tries to establish service semantics using text description and uses it as an additional parameter for service discovery over traditional vector space search to get the best of both methods. We also focus on serving user’s need by automatically composing services whenever required.

3 Proposed Methodology

Figure 1 depicts the overall methodology adopted for the proposed system. We discuss each of these processes in detail below:

3.1 Preprocessing Web Service Description Documents

We propose using a combination of two methods for indexing web service description documents. The first method relies on the exact keywords that define the

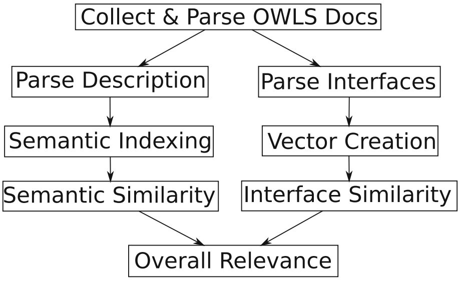


Fig. 1 Proposed methodology

service interface, i.e., input and output. The second method relies on natural language processing techniques to derive the actual functionality provided by the service.

The OWL-S test collection[1] was used as a data set for performing the experiments. We divided the services into seven popular categories like economy, education, communication, food, travel, weapon, and medical. It is difficult to categorize a vast of number of services into these categories strictly and thus we consider a vector space with seven dimensions, each representing a category of web services as listed earlier. Each category has a list of keywords associated with it which is denoted by C_i where $1 <= i <= 7$. We also allow a single keyword to be associated with more than one category.

3.2 *Indexing Web Service Description Documents*

Web service description files typically have tags like <profile:hasInput> and <profile:hasOutput> which specify the respective service interface. We use these interfaces to index the web services. To index a web service, we parse the description document associated with the service to extract the <profile:hasInput> and <profile:hasOutput>. After extracting we extract all the keywords used to describe both the service interface. Let us denote the list of keywords as W_i and W_o. We define two vectors namely V_i (Service Input Vector) and V_o (Service Output Vector). Both the vectors have seven elements (each representing a category) where each element is the number of keywords common to both the category and the service input (for V_i)/output (for V_o). Thus for each service two vectors are created and stored. Mathematically this can be formulated as shown in Eqs. 1 and 2.

$$V_{i_k} = \ | W_i \cap C_k | \ \forall \{k \mid 1 <= k <= 7\} \tag{1}$$

$$V_{o_k} = \ | W_o \cap C_k | \ \forall \{k \mid 1 <= k <= 7\} \tag{2}$$

[1] Available online at http://projects.semwebcentral.org/projects/owls-tc/.

3.3 Measuring Similarity Between Services

The overall functionality of a web service is typically described in a description document associated with web services using human-readable natural language. They tend to give more insight about the actual functionality provided by the web service. Natural language processing techniques can be used to extract the real functionality provided by the web service. We have used a simple word sense disambiguation algorithm to choose the right meaning of all the words in the text description and then establish a context of the web service which can be used for matching the service with users requirements. A context is a set of meanings that represents the functionality provided by a web service.

The algorithm first extracts the text description and then eliminates frequently occurring words such as conjunctions, prepositions, etc., as they do not provide any significant information. Domain modeling can also be used to drop frequently occurring words. After this, all possible synsets of each word are obtained from Wordnet [5]. A context is initialized with all the disambiguated words, i.e., words with single meaning. If no such words are found then a context is established by choosing the most frequently used meaning of a few words having relatively smaller number of synset. The algorithm simply chooses the meaning which is conceptually most similar to the already established context. To compute the similarity between a synset and a context we compute the average similarity of the given synset with each synset in the context as formulated by Eq. 3.

$$sim(syn, context) = \overline{jcn(syn, s_i)} \ \forall \ s_i \ \epsilon \ context \tag{3}$$

where, *syn* is a synset of the given word. To find similarity between synsets we used the JCN Similarity algorithm [4]. After computing similarity of all the synset of a given word with the already established context, we choose the synset with maximum similarity and added it to the context. As such the final context has the set of most similar lemmas. We indexed each web service using the previously established context.

4 Searching Candidate Web Services

With every query user specify the required input, output, and functionality. Based on the two methods of indexing, two different corresponding search methods can be used—vector space based search and semantic search.

In the proposed system, the desired input and output specified by the user are used to construct two search vectors namely S_i and S_o representing the desired interface in the seven-dimensional vector space described earlier. The process of converting the desired interface to corresponding vector is similar to that of converting service interfaces to corresponding vectors. Once we have the search vectors, we search for service vectors similar to search vectors. The similarity between two given vectors is determined using the cosine similarity score, described in Eq. 4.

$$sim(V_1, V_2) = \frac{V_1.V_2}{||V_1|| * ||V_2||} \tag{4}$$

where, V_1 and V_2 are two vectors of same dimension. To find the total similarity between a search query and a service we compute the average similarity of input and output vectors respectively as illustrated in Eq. 5.

$$\text{Total Similarity} = \frac{sim(R_i, S_i) + sim(R_o, S_o)}{2} \tag{5}$$

where, R_i and R_o are the input and output vector of an indexed service. We compute similarity with all the indexed services and then sort the result according to total similarity. After sorting we assign rank to each service.

The semantic search proceeds by iterating over the indexed services and selecting the services with similar context. A context of the user's query is established using the same method as described in previous section. The similarity between two context is computed using the Eq. 6.

$$\text{Semantic Similarity} = \frac{\sum sim(u_i, v_i)}{m \times n} \tag{6}$$

where, u_i ϵ User's Query Context ∀ i = 0, 1, . . . n, v_i ϵ Service's Indexed Context ∀ i = 0, 1, . . . m and *sim*() denotes similarity between 2 given lemmas. The services are sorted according to the context similarity score and each service is assigned a rank. The final rank of a service is computed as the average of rank assigned by each method. We give equal weightage to both the algorithms to compute the final result. However, the weight of each algorithm can be tuned according to the specific requirement.

5 Automatic Service Composition

It is often the case that a single service in the database is unable to satisfy user's query complete. In such cases, we need to find multiple services that can work together in a given sequence so as to provide the required functionality to the user. In essence the services have to be automatically composed. Service composition is performed as follows:

- Searching for suitable web services that can be composed together to act as a single service
- Arranging the different web services in a particular sequence that yields the desired output

- Conversion of data formats so that output of one service matches the input format expected by the next service in the sequence.

Several solutions to this service composition problem have been proposed based on graphical model of web services [3, 6]. In this section, a methodology to search for multiple web services that can be composed together to serve the user's need, using a graph of interconnected web services is discussed.

5.1 Constructing a Service Interface Graph

A DAG (Directed Acyclic Graph) is constructed to model services and the relation between their interfaces. Each node in this graph represents a web service. A node has several incoming and outgoing edges. An edge from node 'A' to node 'B' signifies that the output yielded by service 'A' is similar to the input accepted by service 'B', i.e., service 'A' and 'B' can be composed together. To construct the graph, we first compute the equivalent input and output vector of all the service in the database. Then, we match the output of each service to the input of every other service, i.e., determine the cosine similarity between the output vector of first service and input vector of second service. If the similarity is found to be greater than a predetermined cutoff then the two services are connected via an edge from first one to the second. After the addition of each edge, the graph is checked for cycles. If any cycles are found to exist in the graph, the newly added edge is discarded. Figure 2 illustrates a flow chart showing all the steps involved in creation of the said graph.

The main objective of this process is to model web services such that the service composition problem can be treated as a simple graph traversal problem. Thus, it is essential to have an acyclic graph. The similarity cutoff for adding edges is chosen to be 0.9. A high value is chosen to ensure that there is almost a perfect match between the interfaces. This cutoff can be determined dynamically based on the average similarity of interfaces and several other domain-dependent factors.

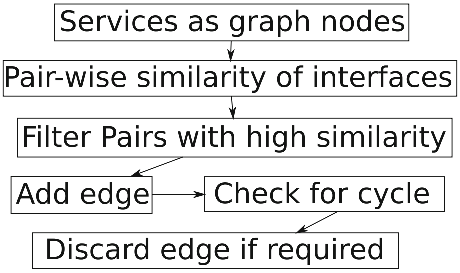


Fig. 2 Creation of service interface graph

5.2 Executing a Service Composition Query

Once the graph is constructed, it has to be traversed for each query. First, an input and output vector corresponding to the user's query is computed. Then a keyword based query as described in previous section is executed. The resultant services are sorted according to the input similarity—and top results (top *k*) are filtered. This gives the top *k* start points of potential composition. For each of these input services, the graph is traversed using the well-known DFS (Depth First Search) Algorithm starting from the input service as source. For every node visited, the similarity between the node's output vector and user's query output vector is determined. Among all the nodes visited, the node with best output vector is selected. The path between the corresponding source node and the node with best output yields the best composition possible for the corresponding source service.

6 Experimental Results and Analysis

The results that we obtain are very encouraging. We are able to obtain very relevant web services given an interface. We have obtained this results with a very small set of keywords in each field (average 40 each). As such, with a large set of keywords in each category we should be able obtain much better results. A few sample Query and their corresponding results have been listed in Table 1. The cosine similarity values obtained for each service in the result set are also mentioned.

In this section, an estimate of the asymptotic run-time complexity analysis of all the major steps involved in our proposed algorithm is presented. The first step is

Table 1 Observed results for some sample queries

Input	Output	Services and Similarity
Car	Price	3wheeledcar_price_service (1.0), car_price_service (1.0), citycity_arrowfigure_service (1.0), lenthu_rentcar_service (0.972)
Missile	Range	missile_lendingrange_service(0.971), missile_givingrange(0.933), ballistic_range_service(0.918)
Location	Distance	sightseeing_service(1.0), DistanceInMiles(0.908), calculate_betwee_Location(0.903), surfing_service(0.901)
Medical	Bed	hospital_investigatingaddress_service(0.789), medicalclinic_service(0.670), SeePatientMedicalRecords_service(0.640)

to parse the description document of a service which is dependent on the length of the description document. Since the length of the description document is roughly same for all services, this step takes constant time. To construct the vectors that can represent the service in 7-D vector space we have to search through all the keywords belonging to all the categories. Thus constructing vectors take $\mathcal{O}(C)$ time, where C is the number of keywords across all categories. Thus the complexity of indexing N services is $\mathcal{O}(NC)$. The next step is construction of DAG. The input vector of each service is matched with output vector of every other service. Thus the complexity of graph construction is $\mathcal{O}(N^2)$, where N is the number of service. The check for cycle formation is linearly dependent on the number of nodes in graph and thus it does not add anything to the complexity. However, the graph is constructed only once when the server is started and thus we can afford it to be slower. Whenever a new service is added to the database, i.e., a new node is appended to the graph, its input and output have to be matched with every other service and thus the complexity of adding new node will be $\mathcal{O}(N)$.

The next step is executing user's query. The complexity of constructing query vector is again $\mathcal{O}(C)$. Finding cosine similarity between query vector and a service takes constant time. Thus the overall complexity of vector space search is $\mathcal{O}(N + C)$. Constructing query context will also take constant time since the number of keywords given by user as a part of their query will typically have a constant upper bound. Again the time complexity of traversing the whole data set and compute context similarity is $\mathcal{O}(N)$. Once we filter top K result, we have to rank them which takes $\mathcal{O}(K \log K)$ time. In a practice, each node will have very few outgoing edges on an average. Thus, assuming the number of edges in the graph is proportional to the number of nodes, the time complexity to execute a service composition query is $\mathcal{O}(N)$.

7 Conclusion and Future Work

In this paper, a novel approach for automatically determining service composition candidates for a given user requirement is presented. The proposed method offers a lot of scope for further improvements. First, we have manually labeled each category with corresponding keywords but the system should be able to learn keywords for each category automatically over time using machine learning techniques. Second, the weightage to results obtained from multiple approaches can be tuned for different domains to obtain better results. Third, domain modeling concepts can be used to improve the word sense disambiguation of the synsets obtained. Certain synsets that have no relevance in a given domain can easily be eliminated by this approach. As the information content of different words is also domain dependent, words frequently encountered in a given domain can be discarded while parsing the description documents. This can help in further optimization and improvement in the performance of the proposed methodology during the process of determining service composition candidates.

References

1. Cuzzocrea, A., Fisichella, M.: Discovering semantic web services via advanced graph-based matching. In: Systems, Man, and Cybernetics (SMC), 2011 IEEE International Conference on. pp. 608–615. IEEE (2011)
2. Fethallah, H., Chikh, A.: Automated retrieval of semantic web services: a matching based on conceptual indexation. Int. Arab J. Inf. Technol. 10(1), 61–66 (2013)
3. Hashemian, S., Mavaddat, F.: A graph-based approach to web services composition. In: Applications and the Internet, 2005. Proceedings. The 2005 Symposium on. pp. 183–189 (Jan 2005)
4. Jiang, J.J., Conrath, D.W.: Semantic similarity based on corpus statistics and lexical taxonomy. CoRR cmp-lg/9709008 (1997), http://arxiv.org/abs/cmp-lg/9709008
5. Miller, G.A.: Wordnet: a lexical database for english. Communications of the ACM 38(11), 39–41 (1995)
6. Oh, S.C., On, B.W., Larson, E., Lee, D.: Bf*: Web services discovery and composition as graph search problem. In: e-Technology, e-Commerce and e-Service, 2005. EEE '05. Proceedings. The 2005 IEEE International Conference on. pp. 784–786 (March 2005)
7. Peng, Y., Wu, C.: Automatic semantic web service discovery based on assignment algorithm. In: 2010 2nd International Conference on Computer Engineering and Technology. vol. 6 (2010)
8. Platzer, C., Dustdar, S.: A vector space search engine for web services. Third European Conference on Web Services (2005)
9. Sangers, J., Frasincar, F., Hogenboom, F., Chepegin, V.: Semantic web service discovery using natural language processing techniques. Expert Systems with Applications 40(11), 4660–4671 (2013), http://www.sciencedirect.com/science/article/pii/S0957417413001279
10. Truong, H.L., Dustdar, S.: A survey on context aware web service systems. International Journal of Web Information Systems 5(1), 5–31 (2009)

A Quality-Centric Scheme for Web Service Ranking Using Fuzzified QoS Parameters

Mandar Shaha and S. Sowmya Kamath

Abstract Service composition, the process of combining already available basic services to provide a new, enhanced functionality, helps in serving diverse user requirements and promotes rapid application deployment. One of the premises for achieving service composition is to consider the quality of service parameters like availability, response times etc., of the constituent services, so that effective ranking can be obtained. However, based on user need, multiple criteria may need to be considered during QoS-based ranking, due to which it may be difficult to provide accurate and precise values with respect to a particular QoS parameter. In this paper, we address this problem by incorporating the theory of fuzzy logic using fuzzy variables. We propose a new scheme that focuses on computing the combined values of various QoS parameters, for enhancing web service recommendation. The proposed scheme has been applied to the real-world datasets, with encouraging results.

Keywords Web service ranking · Fuzzy modeling · Quality of service · Nonfunctional requirements · Web service recommendation

1 Introduction

Owing to the popularity of web services (WS) as the technology of choice for service-oriented architecture (SOA) based system design, there has been growth in the availability of services, many of which provided very similar functionalities. The trend is now towards novel application development based on composite WS. Composite WS are those specialized service workflows, where, basic services providing different but related functional services may be combined to implement a completely

M. Shaha (✉) · S. Sowmya Kamath
Department of Information Technology, National Institute of Technology Karnataka,
Surathkal, Mangalore 575 025, Karnataka, India
e-mail: mandar.shaha90@gmail.com

S. Sowmya Kamath
e-mail: sowmyakamath@nitk.ac.in

P.K. Sa et al. (eds.), *Progress in Intelligent Computing Techniques: Theory, Practice, and Applications*, Advances in Intelligent Systems
and Computing 719, DOI 10.1007/978-981-10-3376-6_42

new, customized service. As there exist various limitations to the simple functionalities it becomes a necessity to combine diverse web services, in view of attaining complex functionalities as required by envisioned business applications [1].

A significant requirement while developing a composite service is to effectively consider both functional and nonfunctional capabilities of their constituent services. The functional capabilities are accessible through its service description in the form of WSDL (Web Service Description Language). The nonfunctional capabilities of a service are expressed in the form of Quality of Service (QoS) parameters. QoS has been generally utilized as a standard approach to assess the nonfunctional parameters of a web service.

QoS parameters considered for WS-based applications include response time, reliability, invocation fee, and security [2]. Hence, it is given that QoS factors play a crucial role in different WS management tasks. QoS also serves as the primary criteria for differentiating between multiple WS that offer similar functionality. As a result, many QoS-aware or QoS-based approaches have been proposed for problems like WS discovery, selection, and composition [3, 4].

One of the crucial limitations in QoS-based approaches, is that it is often counterproductive to merely provide the exact values for a considered QoS parameter. The most intuitive way in which users can indicate their quality preferences is by using relative performance metrics like ‘good service’, ‘fast delivery’, etc. In this paper, we aim to address this problem, by using the concepts of Fuzzy logic. We considered four QoS parameters—*response time* (total time taken to respond to a request for service), *latency* (time elapsed since user generated request till the desired output is displayed), *availability* (the total time a service can be used in a specified unit time interval) and *throughput* (number of requests handled by service per unit time). The proposed approach also uses a scheme for combining two different QoS parameters, viz., *response time*, and *latency*, based on which WS can be effectively ranked.

The outline of the paper is as follows: Sect. 2 presents a discussion on relevant existing work in the area of QoS-based service discovery and recommendation. Section 3 describes the proposed fuzzy logic based WS ranking approach using QoS parameters. Section 4 describes the observed results and the effect of the proposed composite QoS parameters on WS recommendation, followed by conclusion and references.

2 Related Work

WS discovery is an essential task in WS application development and management. As the number of WS grows exponentially, the emphasis is towards composing two or more different services to provide a new functionality, instead of creating another new service for providing new functionality. To combine services, a service provider essentially has to search in a wide range of service collection, and consider many available services providing the same functionality, and finally decide the ones most relevant for composition. For choosing candidate service for composition, a service provider can consider various QoS parameters, as required by the application under development.

Many researchers have worked upon linking statistical approaches for the web service recommendations using nonfunctional parameters. Dai et al. [5] proposed to achieve automation in WS recommendation to service providers to enhance the process of WS discovery. Their focus was on extending a goal-oriented mechanism for WS discovery, using rich functional and nonfunctional parameters, like the service's signature, domain ontology and preconditions, etc. Mishra et al. [6] devised a new system that considered the sequential information in web navigational patterns, with the content information. Their system utilized a technique called the similarity upper approximation and then applied singular value decomposition (SVD) for providing service recommendations to users. In Benaboud et al. [7] work, the functionality match is carried out between services using OWL-S (Web Ontology Language for Services), then functionally similar services are ranked based on the QoS score.

Liu et al. [8] proposed an algorithm called BB4EPS, that combines services based on QoS attributes, where the availability and reliability are considered separately, but here, their consolidated impact is not evaluated. Lin et al. [9] designed a heuristics-based RQSS (Relaxable QoS-based Service Selection) algorithm which uses multidimensional, multi-choice knapsack problem, for composing services. Here, QoS attributes like execution time, reliability, availability, reputation (based on user feedback), and price were used for composing the services.

Gouscos et al. [10] proposed a technique where QoS attributes were classified as static (price, promised response time and failure probability, which are stored in UDDI) and dynamic (actual response time, rate of failure, either stored in WSDL or provided by an information broker). But, they did not address the problem of dealing with out-of-date QoS information, once the data is stored in the UDDI. Huang et al. [11] proposed a single QoS-based service discovery technique, where the service with best QoS attributes is selected and a service with best performance in an entire workflow is selected by a process called QoS-based optimization. To help discovering proper services according to both service provider's and service consumer's preferences and expectations, different multiagent-based approaches have been proposed by Hwang and Chao [12]. In contrast to the approaches discussed, we propose a fuzzy logic based approach that considers both the direct and combined effect of multiple QoS parameters on WS ranking, during WS discovery. In the proposed technique, a composite QoS parameter that includes both response time and latency for ranking WS providing similar functionality.

3 Proposed Methodology

The overall methodology for ranking the web service is shown in Fig. 1. The process involves applying a process of fuzzification to the crisp values of QoS parameters to get the fuzzy sets along with degree with which a crisp value is belonging to a particular fuzzy set. These are given as input to the next phase where inference rules

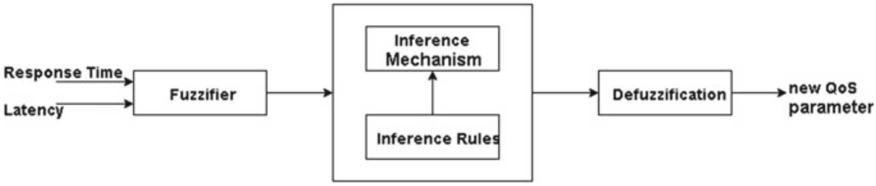


Fig. 1 Fuzzy ranker layout

are applied to generate the fuzzy sets for the proposed composite QoS parameter. Defuzzifying these resultant fuzzy sets will generate the crisp value of the proposed composite QoS parameter, using which the services are ranked.

Fuzzy logic theory basically allows variables to take any real value between 0 and 1. Hence, it is a many-valued logic, in which the concept of partial truth or falsity is valid, unlike Boolean logic. The proposed fuzzy QoS ranker deals with both response time and latency based on a fuzzy rule based decision maker. This component aids in the computation of a new QoS value for all relevant WS according to response time and latency. An overview of this process is shown in Fig. 1. Each of the components is discussed in detail next.

3.1 QoS Fuzzifier

In this stage, the crisp values of QoS parameters considered—response time, latency, availability, throughput, etc.—are taken as input, and then undergo a process of fuzzification. The result is in the form of QoS fuzzy sets and the degree to which the input crisp values belong to these fuzzy sets, which is determined by using membership functions(MF).

The fuzziness in a fuzzy set is determined by its MFs [13] and hence, MFs are the basic building blocks of fuzzy set theory. As the shapes of MFs largerly affect the fuzzy inference system mechanism, shapes of MFs are highly crucial for a specific problem. Different MFs are available—Gaussian Membership Function, Trapezoidal Membership Function, Triangular Membership Function, etc. In the proposed technique, the Trapezoidal MF was chosen, as it allows more number of services to belong to a particular class with higher values, when compared to triangular MF. It also requires less computation when compared to Gaussian MF.

First, MFs are defined for a range of values of a fuzzy set and provide a membership value (from 0 to 1) for each value in that range. This membership value indicates the degree to which the given crisp value belongs to a particular fuzzy set. Next, the range of values of QoS parameters is noted and this range is split into 5 equal parts, with each part corresponding to a fuzzy set. A linguistic variable is assigned to each part, based on its lower limit and upper limit. For example, if the response time range is from 1 to 100 s, it is split into five parts along with linguistic variables shown in Table 1.

Table 1 Fuzzy sets for response time (Nonoverlapping and overlapping)

Fuzzy sets	Nonoverlapping ranges	Overlapping ranges
Very low (VL)	1–20	1–26
Low (L)	20–40	14–46
Medium (M)	40–60	34–66
High (H)	60–80	54–86
Very high (VH)	80–100	74–100

Now, if the graph of membership degree versus response time is plotted with the ranges on the x-axis, the specific degree between 0 and 1 with which each crisp value of response time belonging to a particular fuzzy set (given by the y-axis value) can be found. The ranges shown in column 2 of Table 1 are nonoverlapping and one of the disadvantage of such partitioning is that, it cannot provide proper mapping of degree values to boundary values. For example, if a response time value is 40, it belongs to fuzzy sets *"Low"* and *"Medium"* with a degree 0.0, which is incorrect. Hence, to provide proper justification to boundary values of these parts, it is better to define overlapping ranges, as shown in column 3 of Table 1.

The part range should be in the range of 0 and 100. Trapezoidal MFs are redefined for these new part ranges and each function is represented in the form of 4 points, corresponding to the points of the trapezium. For a given value of response time, it is possible to get the membership degree of that value by plotting a vertical line from that value and finding the y-coordinate where this vertical line cuts one of the edges of a MF. For values falling in overlapping ranges, this vertical line might produce two membership values corresponding to two different fuzzy sets it belongs to. In the same way, fuzzy sets corresponding to the QoS parameter *Latency* are also determined. These fuzzy sets for response time and latency are input to the next stage, *Inference Mechanism*.

3.2 Inference Mechanism

This stage takes as input the fuzzy sets of response time and latency and outputs fuzzy sets of new value of resultant fuzzy set for (RT, L) using Mamdani Style of Inference Mechanism [14]. To explain this inference mechanism, consider the below Inference Rules Matrix given in Tables 2 and 3. The fuzzy sets defined are as follows: VH (very high), H (high), M (medium), L (low) and VL (very low).

As can be seen from Table 2, if fuzzy set of response time is very high and fuzzy set of latency is also very high, then the associated rule indicates that the fuzzy set of new value of combined/composite QoS parameter (RT, L) is very low. This is because, intuitively, it is desirable that a service should have low response time and low latency. But, this approach indicates that the new value of fuzzy set for resultant (RT, L) belongs to very low class.

Table 2 Inference rules for response time and latency

Latency	Response time				
	VH	H	M	L	VL
VH	VL	VL	VL	L	M
H	VL	L	L	M	M
M	L	L	L	M	H
L	M	M	H	H	VH
VL	M	H	H	VH	VH

Table 3 Inference rules for availability and throughput

Availability	Throughput				
	VH	H	M	L	VL
VH	VH	VH	H	H	M
H	VH	H	H	M	M
M	H	M	L	L	L
L	M	M	L	L	VL
VL	M	L	VL	VL	VL

Using this Inference table, each fuzzy set of latency is compared with each fuzzy set of response time to obtain a list of fuzzy sets of resultant (RT, L) value. The membership degree of each of the resultant fuzzy set is the minimum of the membership degree of two fuzzy sets being compared. This list of fuzzy sets and the corresponding membership degree is the input to next step, *Defuzzification*.

3.3 Defuzzification

Defuzzification is the last step in fuzzy inference system. This step takes as input the fuzzy sets of resultant (R,TL) QoS parameters and generates new value in the crisp format. Several different defuzzification methods [15] like maximum membership principle, center of gravity (CoG) technique, weighted average method, etc., are available. The CoG defuzzification method is the most accurate of all and hence was chosen for this problem.

The CoG method is also known as centroid method or center of area (CoA) defuzzification. In this method, an aggregate area is determined corresponding to the resultant fuzzy sets and their membership degrees determined in the previous step. A vertical line dividing this aggregate area into two equal masses is determined and the x-coordinate of this vertical line gives the new value of QoS values in crisp format, which is the required new value. Mathematically, the centroid value is obtained by the summation formula as shown in Eq. 1.

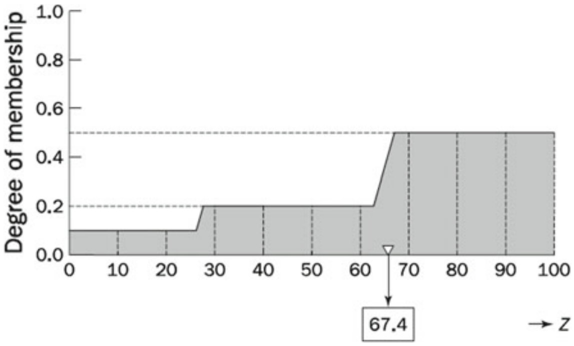


Fig. 2 Defuzzifying the resultant fuzzy sets

$$COG = \frac{\sum_a^b \mu_A(x)xdx}{\sum_a^b \mu_A(x)dx} \tag{1}$$

Figure 2 shows degree of membership with which the resultant variable belongs to each class, and can be used to illustrate how Eq. 1 maps fuzzy sets to crisp values. Here, resultant variable belongs to three classes with a degree of membership 0.1, 0.2, and 0.5, respectively. Now, to defuzzify the value of a resultant variable applying COG formula to the graph in Fig. 2, we get -

$$COG = \frac{(0 + 10 + 20) * 0.1 + (30 + 40 + 50 + 60) * 0.2 + (70 + 80 + 90 + 100) * 0.5}{0.1 + 0.1 + 0.1 + 0.2 + 0.2 + 0.2 + 0.2 + 0.5 + 0.5 + 0.5 + 0.5} \tag{2}$$

which gives the values as 67.4. In this way, the individual fuzzified QoS parameters for response time and Latency, can be mapped to the proposed composite QoS parameter (RT, L), again represented in fuzzy sets. Finally, these fuzzy sets are defuzzified into crisp composite (RT, L) QoS values, which is used for ranking WS composition candidates as per user's requirements.

4 Experimental Results

To evaluate the proposed approach in ranking WS during service discovery, a QWS[1] dataset [16], in which we have considered response time, latency, availability, and throughput QoS parameters for 1064 real-world WS. To illustrate the effectiveness of the proposed approach, we consider a set of eight different web services with the QoS parameter (response time, latency, availability, and throughput) as shown in Table 4. The range of response time is noted and split into five sub-ranges of equal width, with each sub-range representing a fuzzy set. The fuzzy sets (VL, L, M, H and VH) for response time, formed using trapezoidal MF are as shown in Fig. 3. Similarly, fuzzy

[1] Available at http://www.uoguelph.ca/~qmahmoud/qws/.

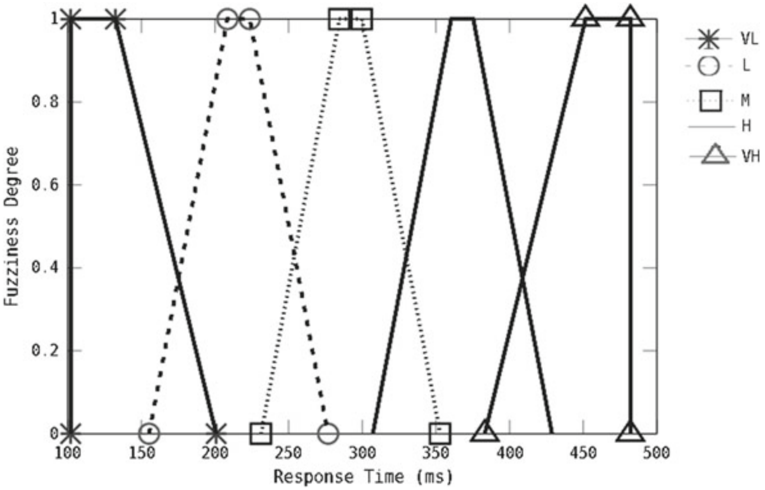


Fig. 3 Fuzzy set ranges for response time

Table 4 QoS metrics for various available web services

Services	Response time(ms)	Latency (ms)	Availability (%)	Throughput (Req/s)
WS1	302.75	187.75	89	7.1
WS2	482	1	85	16
WS3	126.17	22.77	98	12
WS4	107.87	58.33	87	1.9
WS5	107.57	18.21	80	1.7
WS6	255	40.8	98	1.3
WS7	136.71	11.57	76	2.8
WS8	102.61	41.66	91	15.3

sets are identified for the other QoS parameters considered—latency, availability, and throughput also.

For the given crisp values of response time and latency, fuzzy sets are identified such that every crisp value may belong to at most two fuzzy sets. These fuzzy sets along with the degree by which the value belongs to a particular fuzzy set are given as inputs to the inference mechanism. Based on the inference rules, the fuzzy set of the output is determined. Now, using CoG defuzzification technique, a crisp value is obtained for this output fuzzy set, with the help of membership values of input variables. As per the inference rules, the web service with the highest defuzzified value is the best.

To determine the best service, the response time and latency values of each WS are combined to its (RT, L) value and its availability and throughput values are combined to get its (A, TP) value. Table 5 depicts the resultant fuzzified values considering

Table 5 Fuzzy set ranges for QoS parameters

Fuzzy set	Response time (ms)	Latency (ms)	Availability (%)	Throughput (%)
VL	102–200.8	1–49.62	0–26	1–4.9
L	155.2–276.8	27.18–87.02	14–46	3.1–7.9
M	231.2–352.8	64.58–124.42	34–66	6.1–10.9
H	307.2–428.8	101.98–161.82	54–86	9.1–13.9
VH	383.2–482	139.38–188	74–100	12.1–16

Table 6 QoS metrics for various available web services

Web service	(RT, L) fuzzified values	(A, TP) fuzzified values
WS1	273.41	95.17
WS2	386.65	107.30
WS3	606.13	110.33
WS4	608.15	88.31
WS5	613.13	80.97
WS6	537.89	95.65
WS7	603.12	80.97
WS8	613.13	110.33

(RT, L) and (A, TP). For web service WS1, response time is 302.75 which is high and latency is 187.75 which is low. Based on inference rules, if RT is high and Latency is low, the output (RT, L) should be of medium fuzzy set (M). Hence, the corresponding defuzzified crisp (RT, L) value is 273.41 as shown in Table 6. For web service WS8, availability is 91% which is very high and throughput is 15.3 which is also very high. Based on inference rules, if A is very high and TP is very high, the output (A, TP) should be of very high fuzzy set. The corresponding crisp (RT, L) value is 110.33 (as can be seen from Table 6). Hence, higher the fuzzified value, higher is the ranking of a service. Based on the results of the service ranking that can be achieved from the fuzzified values tabulated in Table 6, it can be seen that the proposed fuzzy logic based approach works well.

5 Conclusion and Future Work

In this paper, we proposed a fuzzy logic based technique for WS ranking based on composite QoS parameters, as per user requirement. The aim is to solve the problem of ranking the web services based on different QoS parameters. Normally, when the user is given the choice of selecting web services with same functionality and different QoS parameters like response time and latency, its difficult to choose the best one. In order to obtain an efficient ranking system, a rule-based fuzzy decision approach was presented that deals with both response time and latency. This approach effectively combines both the QoS parameters and computes a new composite QoS parameter, using which functionally similar services can be ranked. As part of future work, we intend to achieve new composite QoS parameters, which can help an application designer in deciding the best service among a pool of functionally similar WS with significantly less time and effort.

References

1. Gabrel, V., M. Manouvrier, and C. Murat. “Web services composition: Complexity and models.” Discrete Applied Mathematics (2014).
2. Barbon, Fabio, et al. “Run-time monitoring of instances and classes of web service compositions.” Web Services, International Conference on. IEEE, 2006.
3. Yan, Hai, Wang Zhijian, and Lu Guiming. “A novel semantic Web service composition algorithm based on QoS ontology.” Computer and Communication Technologies in Agriculture Engineering (CCTAE), IEEE International Conference On, 2010.
4. Alexandre Sawczuk, Hui Ma, and Mengjie Zhang. “A graph-based particle swarm optimisation approach to qos-aware web service composition and selection.” Evolutionary Computation (CEC), 2014 IEEE Congress on. IEEE, 2014.
5. Dai, Bixiang, and Xinke Li. “An enhanced goal-based semantic web service discovery.” Knowledge Acquisition and Modeling, 2009. KAM’09. Second International Symposium on. Vol. 1. IEEE, 2009.
6. Mishra, Rajhans, Pradeep Kumar, and Bharat Bhasker. “A web recommendation system considering sequential information.” Decision Support Systems 75 (2015): 1–10.
7. Rohallah Benaboud, Ramdane Maamri, Zaidi Sahnoun. Towards scalability of reputation and qos based web services discovery using agents and ontologies. In Proceedings of the 13th International Conference on Information Integration and Web-based Applications and Services 2011; ACM; 262269.
8. Min Liu, Mingrui Wang, Weiming Shen, Nan Luo, Junwei Yan. A quality of service (QoS)-aware execution plan selection approach for a service composition process. Future Generation Computer Systems 2012; 28(7); 1080–1089.
9. Chia-Feng Lin, Ruey-Kai Sheu, Yue-Shan Chang, Shyan-Ming Yuan. A relaxable service selection algorithm for QoS-based web service composition. Information and Software Technology 2011; 53(12); 1370–1381.
10. Gouscos, Dimitris, Manolis Kalikakis, and Panagiotis Georgiadis. “An approach to modeling Web service QoS and provision price.” Web Information Systems Engineering Workshops, 2003. Proceedings. Fourth International Conference on. IEEE, 2003.
11. Huang, Angus FM, Ci-Wei Lan, and Stephen JH Yang. “An optimal QoS-based Web service selection scheme.” Information Sciences 179.19 (2009): 3309–3322.
12. Chen, S., Hwang, C., Hwang, F. Fuzzy Multiple Attribute Decision Making: Methods and Applications, Springer-Verlag, Berlin, New York, 292–323, (1992).
13. Zadeh, Lofti A. “Fuzzy logic.” Computer 4 (1988): 83–93.
14. Mamdani, Ebrahim H., and Sedrak Assilian. “An experiment in linguistic synthesis with a fuzzy logic controller.” International journal of man-machine studies 7.1 (1975): 1–13.
15. Negnevitsky, Michael. Artificial intelligence: a guide to intelligent systems. Pearson Education, 2005.
16. Al-Masri, E., and Mahmoud, Q. H., “QoS-based Discovery and Ranking of Web Services”, IEEE 16th International Conference on Computer Communications and Networks (ICCCN), 2007, pp. 529–534.

Enhancing Web Service Discovery Using Meta-heuristic CSO and PCA Based Clustering

Sunaina Kotekar and S. Sowmya Kamath

Abstract Web service discovery is one of the crucial tasks in service-oriented applications and workflows. For a targeted objective to be achieved, it is still challenging to identify all appropriate services from a repository containing diverse service collections. To identify the most suitable services, it is necessary to capture service-specific terms that comply with its natural language documentation. Clustering available Web services as per their domain, based on functional similarities would enhance a service search engine's ability to recommend relevant services. In this paper, we propose a novel approach for automatically categorizing the Web services available in a repository into functionally similar groups. Our proposed approach is based on the Meta-heuristic Cat Swarm Optimization (CSO) Algorithm, further optimized by Principle Component Analysis (PCA) dimension reduction technique. Results obtained by experiments show that the proposed approach was useful and enhanced the service discovery process, when compared to traditional approaches.

Keywords Web service discovery · Bio-inspired algorithms · Document clustering · Semantics · Swarm intelligence

1 Introduction

Service-oriented computing (SOC) is the computing model that uses services as basic components for application development. To build a service-centric application model, SOC extends a service oriented architecture (SOA), as a means for reorganizing software applications and frameworks into a set of collaborative services. Web services are currently the most popular model for implementing a SOA system, as

S. Kotekar (✉) · S. Sowmya Kamath
Department of Information Technology, National Institute of Technology Karnataka,
Surathkal, Mangalore 575 025, Karnataka, India
e-mail: sunainakotekar@gmail.com

S. Sowmya Kamath
e-mail: sowmyakamath@nitk.ac.in

P.K. Sa et al. (eds.), *Progress in Intelligent Computing Techniques: Theory, Practice, and Applications*, Advances in Intelligent Systems
and Computing 719, DOI 10.1007/978-981-10-3376-6_43

they are standards based, and employ XML-based protocols for messaging, service description, and data transfer. The capabilities of a Web service are encapsulated in its service description, in the form of a WSDL (Web Service Description Language) document.

The task of searching for relevant Web services for a given requirement is traditionally based on the service name and natural language description [1]. Hence, this search is similar to that performed by search engines and is primarily keyword based. However, several studies have revealed that about 30% of services have <10 words of natural language documentation, while more than 55% of published services do not have any [2, 3]. Also, basic keyword search overlooks the actual functionality of a Web service, by concentrating only on the service name and documentation (if available).

To overcome these drawbacks, intelligent techniques to automatically capture both the functionality of a service and also its functional domain are crucial. This has to be done with available data, instead of relying on service providers to make such information available explicitly. Since a service's WSDL effectively describes its capabilities and also provide information for a client on how to use the Web service, the WSDL document is used as an effective source for determining the functionality of a service. Text mining techniques can be applied to WSDL documents to identify useful components, which describe the actual functionality of the corresponding Web service. Using this functionality-related information, WSDL documents can be clustered to capture their domain, thus achieving search space reduction during the process of service discovery.

Document clustering is the process of categorization of documents into groups of similar documents, where each group represents a particular domain. The process is performed such that intra-group document distance is to be kept low, while inter-group document distance should be high. A distance measure technique is thus the heart of document clustering. Document clustering significantly reduces the search space/domain when it comes to searching a related document(s) with some keywords. Semantics-based techniques that capture the morphological variants of a user provided keyword, can further enhance the retrieval process.

This paper presents an Web service clustering approach based on Cat Swarm Optimization, that emulates the social behavior of cats in nature. The clustering accuracy of the CSO algorithm was analyzed with a traditional algorithm like the K-means basic clustering algorithm. Also, PCA was used for further optimizing clustering purity. The rest of the paper is presented as follows: Sect. 2 discusses existing techniques applied to the problem of clustering Web services to enhance service discovery. Section 3 describes the proposed CSO-based methodology and the proposed enhancements. Section 4 discusses the experimental results and analysis, followed by conclusion and references.

2 Related Work

Discovering Web services based on the functional requirements is need of the hour in search engines. To boost precision in search results, several approaches have focused on clustering Web services based on their functional similarity. Elgazzar et al. [4] suggested a technique for finding similarity between Web services, using the service's WSDL document which describes a particular Web service in detail. Using this computed similarity between each WSDL document corresponding to Web service, the services were clustered using QT clustering algorithm. Nayak et al. [5] explained strategy for finding the affinity between Web services on the basis of Jaccard Coefficient to cluster similar document for the ease of discovery of services. Liu and Wong [6] gave a idea for Web service clustering on the basis of extracted content from WSDL document such as content, hostname, context, and service name.

The drawback of these approaches is that, a traditional algorithm like QT clustering can result in too many clusters of small size. Initially, some clusters of big size will be formed and then the remaining data will get clustered into smaller clusters and all remaining dissimilar service will form a cluster, because of this, purity of the clustering will be very low. Also, creating a similarity matrix of size *nxn* for all *n* WSDL documents is time consuming.

Chu and Tsai [7], first discussed the concept of computational artificial life or computational intelligence, which primarily emulate animal behavior in nature for solving computationally hard problems. These algorithms are usually employed as optimization techniques, and several swarm intelligence based methods that simulate the intelligent behavior of animals are currently available. Particle Swarm Optimization [8] (PSO), Ant Colony optimization (ACO) [9], and also a recent development in the form of Cat Swarm optimization (CSO) [7, 10] that makes use of the social herding behavior of cats in nature. Some problems which are addressed using these nature-inspired algorithm are scheduling, Vehicle routing problem (VRP), Shortest Path problem, Traveling salesman problem (TSP) and data mining, particularly clustering problem, etc. Chu et al. [10] proposed the CSO algorithmic rule that extends two sub-models based on the two most important behavioral character of cats, the "seeking mode" and the "tracing mode". Santosa [11] recommended a method for clustering records in a standard dataset according to their classes using CSO clustering algorithm, which was found to be better than several other clustering techniques.

Our approach is to apply the CSO-based optimization to traditional clustering algorithm like K-means, using the computed functional similarity of service documents. Natural language processing methods like tf-idf (term frequency—inverse document frequency) and morphological analysis are used to obtain the similarity and dissimilarity between service documents. This similarity value is used to effectively cluster service documents into functionally similar groups.

3 Proposed System

The proposed methodology is aimed at describing the problem of extracting the functional information of services, and using this to automatically categorize a set of Web services in a domain specific manner. Figure 1 shows the workflow of the proposed work.

A WSDL document is an inherent source of the functional details of a Web service, so these are first preprocessed. All the natural language terms in the WSDL document are considered as a feature list and extracted by a process called content extraction. Filters like stopword removal, stemming are applied in preprocessing. The outcome of this step is set of keywords from each WSDL document, along with their frequency in dictionary format. This dictionary of words from all documents combined are taken as *attributes*. Next, the tf-idf value is calculated using the Eq. (1).

$$tf - idf_{i,j} = tf_{i,j} * idf_i \tag{1}$$

where i represents the ith document and j represent jth word in the attribute list $tf_{i,j}$ is frequency of attribute j appears in document i as shown in Eq. (2). where as idf_j is calculated using Eq. (3) explains how important an attribute is.

$$tf_{i,j} = \frac{number\ of\ times\ attribute_j\ in\ document_i}{Total\ number\ of\ words\ in\ document_i} \tag{2}$$

$$idf_j = log\frac{N}{|d \epsilon D\ :\ j \epsilon d|} \tag{3}$$

where N is number of documents. idf_j is calculated as logarithmic fraction of total number of documents to number of documents containing jth attribute.

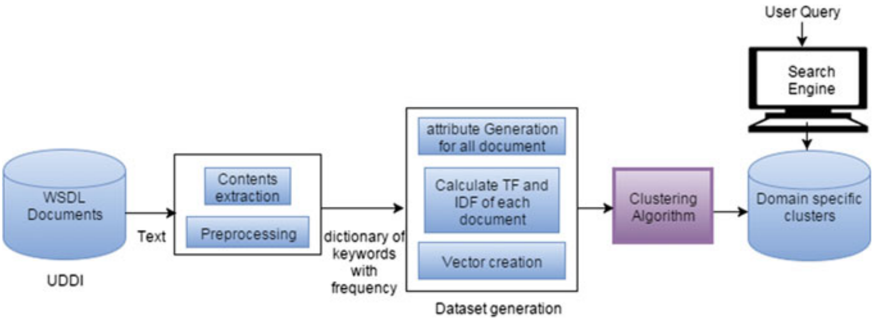


Fig. 1 Proposed methodology

Based on the calculated values of $tf - idf$, the nearness and dissimilarity between the documents are calculated using the Euclidean distance formula as given by Eq. (4).

$$d(x, y) = ||x - y||^2 = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{4}$$

Using this value, the K-means algorithm is applied to the service documents. Here, the number of clusters (K) into which the service documents are to clustered is specified. During the first iteration, K documents are randomly chosen as cluster centers. Based on the Euclidean distance formula given in Eq. (4), the distance between each document and the center chosen is computed and the document if assigned to the nearest center. The calculation of mean of each cluster leads to the discovery of new centroid of the cluster and the process of clustering is continued. Finally, the clusters formed are stored in the repository. Whenever a user searches for a Web service we can find the relevant cluster of WSDL documents to provide them a better result. Each cluster can be tagged with keywords for ease of search. After finding relevant clusters for query, WSDL documents in clusters can be sorted based on relevance.

3.1 Cat Swarm Optimization Based Clustering

To further improve the clustering obtained after K-means clustering algorithm, the CSO technique was applied to the service documents. The standard CSO algorithm defines two sub-models based on real-world cat behavior while hunting for food, called “seeking mode” and “tracing mode”. In CSO, the number of cats to be used in each iteration has to be decided initially. Every cat is described by a feature vector of D dimensions that represent its position, velocities pertaining to every dimension, a dependency of fitness operator (in terms of degree) known by the fitness worth of each cat, and Boolean flag to determine cats’ mode (seeking or tracing). Algorithm 1 illustrates CSO clustering applied to given service documents. The two sub-modules of Algorithm 1 are explained in detail in procedures.

(a) Seeking Mode Four fundamental pillars of seeking mode are: the selected dimension (SRD), seeking memory pool (SMP), self-position consideration (SPC), seeking range of counts of dimension to change (CDC) (shown in procedure: seeking mode). The SSE is computed using Eq. (5) and the distance for reassigning the clusters is computed as per Eq. (4). The seeking mode imitates the process of cats moving around while they are near a prey. So, in CSO, the cluster centers are changed slightly after each iteration, after which the SSE is computed and updated to the smallest value.

$$SSE = \sum_{i=1}^{k} \sum_{x \in D_i} (||x - m_i||^2) \tag{5}$$

Algorithm 1: CSO Algorithm for Clustering

```
Data: Dataset for clustering, number of cluster K, Number of copy
Result: Getting clusters of data
Randomly choose K data as cluster centers $c_1, c_2, ..., c_K$;
repeat
    Enter Seeking mode;
    if $seeking\ SSE < SSE_i$ then
        $SSE_{i+1}$=new SSE
    else
        $SSE_{i+1} = SSE_i$
    Enter Tracing mode;
    if $seeking\ SSE < SSE_i$ then
        $SSE_{i+1}$=new SSE
    else
        $SSE_{i+1} = SSE_i$
until Records are not stable in cluster;
return List of records in different cluster
```

Procedure Seeking mode

```
Data: Parameters SPM, SRD, SPC
Result: Getting intermediate clusters of data
for i=0 to k do
    create SMP times copy of cluster $center_i$ position
    Determine i value
    Compute shifting parameter ($clustercenter_i * SRD$)
    for x=1 to SMP do
        At random add or subtract cluster centroids with the shifting parameter ((SMP*k)
        cluster candidates are obtained at this step)
        Compute distance
        Group data into clusters based on distance calculated
        Compute SSE
        Determine the potential candidate to be recognized as new cluster center by roulette
        wheel selection
return List of records in different intermediate cluster
```

(b) Tracing Mode The tracing mode imitates the behavior of real-world cats, when they intend to attack their prey, by jumping on the prey. So, in CSO, the cluster centers with least SSE are considered and the *velocity* is calculated and updated using Eq. (6). Using this computed velocity, the position of the cat is updated using Eq. (7). For the new position, the SSE is calculated again as before, and we check if the new value is the lowest. If the SSE value is lower than earlier value, then the new position is the updated cluster center.

$$V_{k,d} = v_{k,d} + r_1 * c_1 * (x_{best,d} - x_{k,d}) \tag{6}$$

$$x_{k,d} = x_{k,d} + v_{k,d} \tag{7}$$

Procedure Tracing mode

```
Data: Intermediate clusters and centers
Result: Getting intermediate clusters of data
for i=1 to k do
    Modify $velocity_i$
    Modify $position_i$
    Find new cluster $center_i$
Compute distance
Assign data points into clusters based on distance calculated
Compute SSE
return List of records in different intermediate cluster
```

3.2 Dimensionality Reduction Using PCA

In the proposed clustering methodology, a large WSDL dataset is used, due to which the constructed tf-idf matrix results in large number of attributes. In this tf-idf matrix, most of the values are zero, as a particular term may be relevant only to a few services in the same domain, due to which the tf-idf matrix is highly sparse. To reduce this sparseness, dimensionality reduction technique is used.

Principal Component Analysis (PCA) is a technique used for feature reduction, which involves mapping of data from high-dimensional space to low-dimensional space. To adopt this technique, we compute covariance matrix for the data, using which the eigenvalues are calculated. Based on eigenvectors obtained from largest eigenvalues, we reconstruct the new data matrix with lower dimension. Eigenvectors are called as principal components.

3.3 Similarity-Based WSDL Dataset Generation

When the tf-idf matrix is generated, only the existence of the attribute and its frequency in that particular document is considered. To consider the semantic similarity between terms used in functionally similar services, the morphological variants and synonyms of terms must also be considered. For example, car, vehicle, four wheeler, etc., are related words, while tf-idf would consider these are completely different words. We incorporate a word similarity measure to determine functionally similar documents inside a cluster. To calculate the data matrix, we modify tf-idf equation as shown in (8).

$$tf - idf_{i,j} = tf_{i,j} * idf_j * MaxSimilarity_{i,j} \tag{8}$$

where $MaxSimilarity_{i,j}$ is 1 if the *ith* attribute is present in *jth* document. Else, it is maximum of all similarity found between words in *ith* document to *jth* attribute as

Table 1 Purity of clusters for different datasets

Dataset name	No of records	Attributes	Classes	K-means purity (%)	CSO purity (%)
Iris	150	4	3	67	90
Glass	214	9	6	54	58
Balance scale	625	4	3	61	78
Soybean small	47	35	4	79	83
Wine	178	13	3	70	72
WSDL Documents	1083	644	9	56	63

shown in Eq. (9). Here *n* being total number of words in *ith* document. Similarity is computed using Wordnet [12].

$$MaxSimilarity_{i,j} = \max_{1 \le k \le n} Similarity(word_k, attribute_j) \tag{9}$$

4 Experimental Results

To calculate effectiveness of the CSO-based clustering, several standard datasets (like Iris, Glass, Balance Scale, Soybean Small, and Wine[1]) as well as on the WSDL dataset[2] were taken and experimented upon. The purity or accuracy of clustering is calculated using Eq. (10), where *j* is number of classes and *k* is number of clusters.

$$Purity(\%) = \frac{\sum_0^k \max_0^j (Documents\ belonging\ to\ each\ class)}{Total\ number\ of\ documents} \tag{10}$$

Table 1 tabulates the accuracy of clustering using K-means and CSO on the various standard datasets and on the WSDL dataset. There were three classes of Iris flowers available in the given standard data set, namely “Iris-versicolor”, “Iris-setosa”, and “Iris-virginica”. In the case of WSDL documents, the domain to which a Web service belongs to is taken into account for calculating the purity. Domains are: “education”, “travel”, “food”, “geography”, “medical”, “economy”, “weapon”, “communication”, and “simulation”. This is obtained from the folder hierarchy of the OWL-S TC dataset.

As can be seen from the results, purity level varies with the number of attributes or number of records. It was observed that purity level almost inversely proportional to number of attributes and records collectively. Figure 2 shows comparative analysis of cluster purity for K-means and CSO on different data sets. Table 2 shows the different

[1] Available at https://archive.ics.uci.edu/ml/datasets.html.

[2] Available at http://projects.semwebcentral.org/projects/owls-tc/.

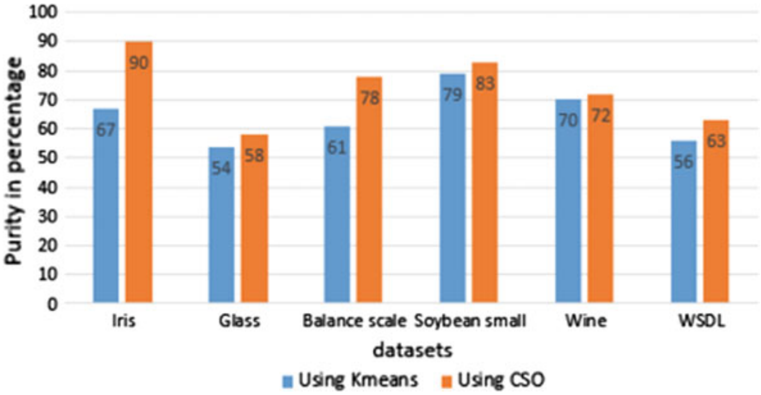


Fig. 2 Observed purity before PCA—K-means versus CSO

Table 2 Purity of clusters for different datasets after applying PCA

Dataset name	No of records	Attributes	Classes	K-means purity (%)	CSO purity (%)
Wine	178	4	3	83	89
WSDL Documents	1083	4	9	60	69

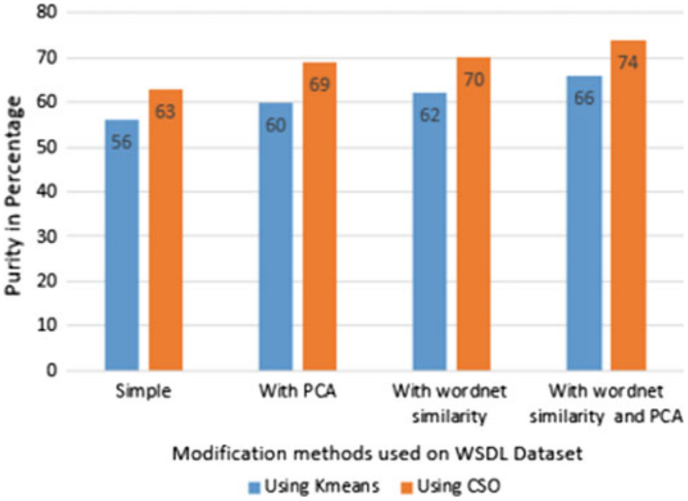


Fig. 3 Observed purity using modified WSDL dataset—K-means versus CSO

results based on data sets after the PCA dimensionality reduction has been applied to the sparse data. The number of attributes are more in the case of Wine and WSDL dataset, and also the data is sparse, so PCA was applied for reducing number of attributes. As a results, clustering purity increased from 63 to 69% and also the run time of algorithm reduced greatly due to less dimension of data.

Figure 3 shows the result comparative analysis graph between simple WSDL dataset and modifying WSDL dataset using PCA, modified WSDL dataset using Wordnet Similarity, modified WSDL dataset after applying PCA and Wordnet Similarity, it was observed that purity of clusters increased when the similarity factor is added to the dataset.

5 Conclusion and Future Work

In this paper, a modified CSO-based algorithm for clustering for Web service was discussed. Text mining techniques were applied to a real-world service dataset to extract their functional information and CSO was applied to these to determine similar groups. The clustering accuracy of CSO was analyzed with traditional K-means basic clustering algorithm. It was found that the purity of clusters obtained by CSO was better than K-means clustering algorithm by about 7%. When PCA-based dimension reduction was applied, the purity of clustering increased to 69% from 63%. This is because there is randomness involved in K-means initial center selection and the algorithm terminates when the centers are stable. But in the case of CSO tracing mode, random change of centers are carried out for determining if potentially better clustering is possible. As part of future work, this purity can be further enhanced by making use of feature selection techniques, along with feature reduction techniques, to capture the best features based on which clustering can be performed.

References

1. Bachlechner, Daniel, et al. "Web service discovery-a reality check". 3rd European Semantic Web Conference. Vol. 308. 2006.
2. Fan, J. and Kambhampati, S., 2005. A snapshot of public web services. ACM SIGMOD Record, 34(1), pp.24–32.
3. Kim, Su Myeon, and Marcel-Catalin Rosu. "A survey of public web services." E-commerce and web technologies. Springer Berlin Heidelberg, 2004. 96–105.
4. Khalid Elgazzar, Ahmed E. Hassan, Patrick Martin, Clustering WSDL Documents to Bootstrap the Discovery of Web Services, 8th IEEE International Conference on Web Services (ICWS'10), Miami, Florida, USA, pp. 147–154, July 2010.
5. Richi Nayak, Data mining in Web services discovery and monitoring, International Journal of Web Services Research, Vol. 5, No. 1, pp. 63–81, January, 2008.
6. Wei Liu, Wilson Wong, Web service clustering using text mining techniques, International Journal of Agent Oriented Software Engineering, Vol. 3, No. 1, pp. 6–26, 2009.
7. S. C. Chu, and P. W. Tsai, Computational Intelligence Based on the Behaviour of Cat, International Journal of Innovative Computing, Information and Control, 3 (1), 2007, pp.163–173.
8. A. Abraham, S. Das, S. Roy: Swarm Intelligence Algorithms for Data Clustering. Soft Computing for Knowledge Discovery and Data Mining, 2008: 279–313.
9. Shelokar, P. S., Valadi K. Jayaraman, and Bhaskar D. Kulkarni. "An ant colony approach for clustering." Analytica Chimica Acta 509.2 (2004): 187–195.
10. S. C. Chu, P. W. Tsai, and J. S. Pan, Cat Swarm Optimization, LNAI 4099, 3 (1), Berlin Heidelberg: Springer-Verlag, 2006, pp. 854858.

11. Santosa, Budi, and Mirsa Kencana Ningrum. “Cat swarm optimization for clustering”. Soft Computing and Pattern Recognition, 2009. SOCPAR’09. International Conference of. IEEE, 2009.
12. Miller, George A. “WordNet: a lexical database for English”. Communications of the ACM 38.11 (1995): 39–41.

Advancement in Personalized Web Search Engine with Customized Privacy Protection

Jeena Mariam Saji, Kalyani Bhongle, Sharayu Mahajan, Soumya Shrivastava and Ashwini Jarali

Abstract Technologies are blooming, needs are growing, larger user data is getting aggregated, and thus privacy becomes a matter of concern in this fast paced, technology driven environment. People are relying mostly on Internet for almost everything they work on or experience. The web search engines confuse us sometimes by giving mixed results. Different people may have variant requirements, and search engines provide same results for same queries, but to different people. In this paper, we intend to solve this problem by a technique of generating online user profiles before firing any query. This user profile would store the user details and the search engine would display results according to this generated profile. We use collaborative filtering and ranking function to filter out the pages according to the preferences of user. We intend to add a feature in our system where in, the users will get a chance to handle their degree of privacy. We offer them two friendly buttons—"Private" and "Public". These buttons will decide whether the user wants to share his details with other users or not. A combination of personalization and privacy would surely be worth a good use for the Internet seekers.

J.M. Saji (✉) · K. Bhongle · S. Mahajan · S. Shrivastava · A. Jarali
Department of Computer Engineering, International Institute
of Information Technology, Pune, India
e-mail: jeenamariamsaji@gmail.com

K. Bhongle
e-mail: bhonglekalyani@gmail.com

S. Mahajan
e-mail: sharau.mahajan@gmail.com

S. Shrivastava
e-mail: soumya.300195@gmail.com

A. Jarali
e-mail: ashusleek1@gmail.com

P.K. Sa et al. (eds.), *Progress in Intelligent Computing Techniques: Theory, Practice, and Applications*, Advances in Intelligent Systems
and Computing 719, DOI 10.1007/978-981-10-3376-6_44

Keywords Customize privacy • Filtering • Generalized profile • Personalized search results • Re-ranking

1 Introduction

Search engines have become a key element for discovering information over the internet. For every problem, we think of internet as a savior. It is often observed that same set of results are displayed to different users for the same query. For example, a doctor wants to search information related to a human face. When he types "face" and hits the search button, he may get Facebook as a search result instead of a human face. Thus in current system, we get mixed results, not the results according to our preferences. Hence it becomes difficult to find for the desired information at one single glance. We often need to go through several other pages in order to find the specific link of information. Sometimes, the results displayed by the search engine may be relevant to the keyword entered by the user, but may not be able to fulfill user's expectations of information need. A user enters the query for which he wants to search information in the search textbox and receives a long list of results or links in lieu of the query entered. The challenge of the search engine is to translate user's simple queries into list of documents that satisfy the different information needs. To overcome this challenge, people came into a conclusion of working with the personalization of search engines.

The profile-based personalized search engine takes the user input, displays the list of results, and also uses the additional information about the user to aid in checking the relevance of the pages. There are various approaches to provide personalization to web search engines. Some of the features determining such approaches are user details, user level interaction, and information which is stored and algorithm which is used to retrieve user details into the search.

The main feature of this paper is that we allow users to control their degree of privacy protection by providing them a Private and Public option in our search engine. These options help the user in deciding their level of privacy according to their requirements. If a user wants to share the browsing queries, he just has to hit the "Public" button and thus he can maintain the transparency accordingly.

In general, our main aim is to develop such a search engine which is privacy protected as well as customized privacy web search engine. The paper is further elaborated into II. Literature Survey and III. Proposed System briefly explaining the purpose of the paper.

2 Literature Survey

Personalization is being accepted by a large set of users to ease the use of web search engines. But despite being proposing it for many years it is still difficult to analyze whether personalization has an adverse effect on all kinds of queries and for different users or not. Dou [1] gives an overview of different problems faced in personalization along with their solutions to it. It is followed that the queries entered by different users often produce the same results altogether, in spite of being variant information goal. A framework based on query logs is developed to ensure massive scale enhancement of personalized search. It is revealed that queries with large click entropy have severe improvement over the common web search. It is seen that different queries has different effectiveness so it is advised that not all the queries should be personalized. The profile-based search strategy mentioned by Z. Dou is not as reliable or effective as the click-based search strategy. It is realized that short-term, long-term contexts and logs are necessary to be analysed for a profile-based search strategy. Thus it is concluded that a combination of both would be reliable.

Deng and Lee [2] presented the personalization of web search engine where the results are displayed according to the preferences set by the users. They introduce us to another mining technique, called Spy Naïve Bayes (NB) which states that the clicked items imply user's choices. It is often seen that for same queries issued by different users, produce same result. However, different users may tend to have different choices for searching a particular query. For resolving this problem of search engine transformation, some research issues are considered. The primary research issue is preference mining which deals with the preference of users of search results from click-through data. Another issue is ranking function optimization which helps in optimizing the retrieval of results according to the user's preferences in search engines.

In the new SpyNB approach, a list of preferences is generated and is fetched by the Ranking Support Vector Machine (RSVM) for optimizing the ranking function for the user. The SpyNB algorithm helps in generating preference fragment pairs used for ranking function. The fragment pairs offer an effective element in making this approach more reliable. Thus it is concluded that SpyNB approach is more productive and flexible than the algorithm existing currently.

Alexander Pretschner and Susan Gauch [3] in paper represented a system where they have explored the ways of incorporating user's interest to the search process in order to enhance the search results. They suggested the method to generate a user profile depending on the way the user would surf the online pages. Combination of three major metrics, time, subject discriminator, and length were used to analyse the user behaviour and create his/her profile accordingly. Here time denotes the amount of time a user spends on a given page, while length refers to the number of characters in the page. On the basis of the analysis done by them, the profiles reflected the user's interest quite well and could be used to deploy more effective information retrieval and filtering. So basically this paper provided a solution to retrieve more relevant search results by using the profiles created on the basis of surfing history of a user.

In order to overcome the issues related to privacy concern from user's perspective, Krause and Horvitz [4] explored and introduced a study of privacy in personalization, where user has an option to share his/her personal information, in return for expected enhancement in the retrieval of more relevant search result. Krause and Horvitz illustrated the methodology based on the graphical analysis survey of the log that saved user's search history. Through this survey they seek to comprehend the utility of personalization that can be actualized by using user's log-based information to analyse his/her willingness to trade the sharing of their personal data with any online services that they are exposed to. Thus they focused mainly on achieving efficient personalized search service using minimum user information.

Lidan Shou and Chen [5] presented a study paper where they used a user-side privacy protection system which is called UPS for personalized web search. They proposed two algorithms, namely Greedy DP (Discrimination power) and Greedy IL (Information Loss) that was used to generalize the user profile in order to avoid exposure of user's personal information while using the search engine. Privacy risk and utility of personalization were the two major predictive metrics used in their proposed algorithm. Also their experimental results acknowledge that the UPS framework could outperform the existing web framework and provide more effective and efficient solution. For future work they suggested to use the better predictive metrics to improvise the performance of the UPS framework.

Xiao and Tao [6] believed that the existing methodology focus on a universal approach that endeavor the same amount of preservation for all persons, without catering for their actual needs. Motivated by this, they came up with the alternative solution of generalizing the whole web search framework based on the concept of personalized k-anonymity. The method has been explained with the help of careful theoretical study of the user information which is used for the research purpose. They have used QI generalization to generalize the various attributes that are taken as an individual's detail information. In this technique if the user provides n number of attributes in a detail table, then after generalization over these tuples/attributes, only n and k detail would be exposed to the outsider hence preventing any kind of information loss. Here k is the tuples that has been generalized and eliminated from the final set of data.

Also the paper has clearly mentioned the drawback of their proposed solution for providing privacy protection in the personalized system and also focuses on developing more optimal alternative generalization strategies.

3 Proposed System

In this proposed system, we are using various technologies to develop a client side privacy protected personalized web search engine. By calling it a client side protected search engine, we mean that the user will have a control over sharing his profile and browsed logs with other users. In this system, we are using a technique known as collaborative filtering where information and patterns involving collaboration among multiple agents are filtered which lets us know the preference of the users. Using collaborative filtering, ranking and rating will be done on web documents. Our main aim is to generate results according to user's preference and lower the risk of disclosing user's sensitive information. Technologies used in this technique are web crawling, web mining, pattern recognition, and application program interfaces (API's).

In our system, initially the user will create an account on search engine. By creating an account the user will create a profile which will be stored in database server. Privacy is also provided to individual profiles. While creating a profile, the user will provide personal details like address, profession, interests, hobbies, etc. After signing up, the user will login into the respective account and will start browsing by issuing a query. We need a client database server which would take the responsibility of storing the user profiles. Generalization of the profiles [5] would take place alongside and will be sent to the central server. Let us assume these generalized profiles as "G". Generalized Profile (G) will be sent to the web crawler and then the functioning will begin.

When the query is issued, it is first preprocessed and then sent to the World Wide Web where the web crawlers analyse the entered query and crawl to different web

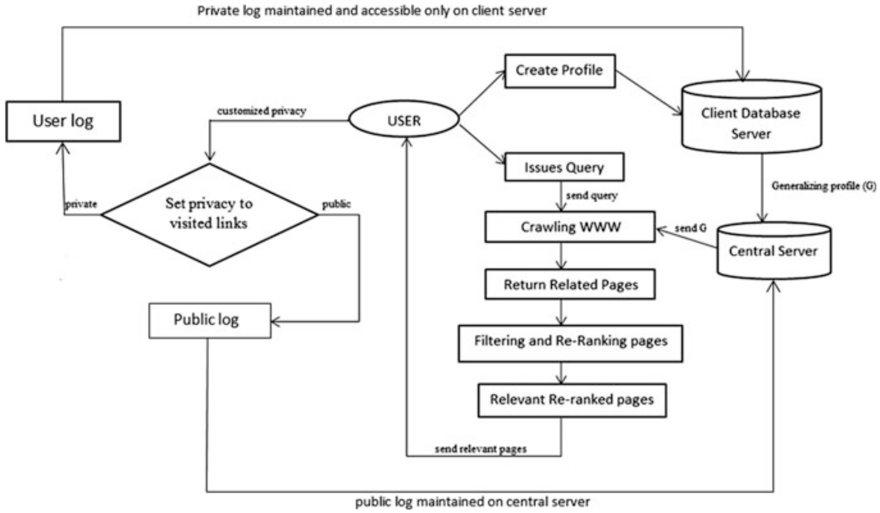


Fig. 1 System flow

pages to collect information from different sources. The web crawler then returns the pages related to the issued query. Since we are using collaborative filtering in our system, the related pages will get filtered. After filtering the crawled pages, ranking function is performed on them. And that is how relevant results are sent back to user. The user-system interaction will be possible because of an API in the middle layer. Figure 1 shows the brief functioning of our system.

In our system, we are using Spy Naïve Bayes algorithm [2] and Deep Search algorithm for reverse searching of relevant document. Privacy is provided to users by providing them with two buttons—“Private” and “Public” respectively. When the “Private” button is clicked, the link being customized to private status will be maintained in the private log, which will be stored on the client server and accessible only to the user. Whereas when “Public” button is clicked, the links with public status will be maintained in public log that is stored in server and will be accessible to all. Thus in this way we combine privacy with personalization in our system making it users decision whether to share his details and visited sites with server which in turn uses the public log for business and similar purposes making this public log consisting user history accessible to other stakeholders like advertisers, researchers, analysts and similar third party member.

3.1 Basic Steps Involved in Personalized Web Search (PWS)

Let us consider the following set of tuples:

P = {Set of user profiles}
Q = {Set of queries given by a user}
R = {Set of Response/search results given back to user}
G = Generalized profile for every P
N = Number of overall results related to Q

- Creating generalized user profile

1. Initially a user will register with the PWS engine by creating his/her own profile P which will consist of attributes like his/her name, gender, age, profession, interest, and other such related personal detail.
2. This profile P will be processed in such a way, that only the attributes which are required for the further processing will be collaborated together to form a new generalized profile G. This is done in order to avoid any kind of user’s personal information loss.

The information related to the user profile will be saved on the client-side itself for reducing the privacy concern. While the updated generalized profile G can be saved on the serve-side, since it would be required while filtering the relevant search results as per the user’s profession and interest. Storing G on the server side assures the better response time for processing at the same time reduces the complexity while filtering and ranking the pages as the need of communicating with the client server would be avoided.

- Ranking and providing customized privacy

1. User can further browse a query Q in the search box which in return will send this Q to the main server where the crawling over World Wide Web would be initialized.
2. The process of crawling will give N results related to the query Q send by the user. These N results would be filtered and re-ranked based on the generalized user profile G.

Suppose R[n] is the set of result related to Q returned after crawling, 'i' is the index for every individual page/link in R[n] and 'rank' is the ranking of the page,

Then if R[i] == G && rank == high
Then set R[i] first

Repeat till whole result set R[n] is sorted and re-ranked based on G.

3. Once the filtering is done the user response would be created with completely new set of re-ranked page results R as per the user's interest.
4. For every Q the user would be given the choice to set his status either private or public. The status here specifies whether the user want to share the visited links with the other users.

Here the other users are the stakeholders consisting both the registered PWS users and advertisers.

This way the user would be ensured that his pattern of going through the result set is not being intruded by others and neither any of his personal information is exposed to the outsider.

3.2 Tools and Technologies Used

Tools required in our system includes Windows 2007 or above, JDK 1.7 and Tomcat Apache 7.0 and MySQL database. For our system to be executed, we need minimum hardware which would include Processor Pentium 4 or above and minimum hard disk space of 2 GB.To communicate and get connected we need some hardware interfaces like Ethernet, modem and Wi-Fi router, as well as some software interfaces like web browsers, DB2, Eclipse, servlets, AJAX, JSP, and Operating System.

We make use of communication interfaces like Internet, Web Server, and HTTP protocol basically on the central repository.

3.3 Objectives of Proposing the System

- To provide relevant search results based on the users choice.
- To ensure privacy protection to the user's personal information.
- To simplify the filtering process using simple sorting based solution.
- To provide customized privacy service to the user for sharing his/her visited results and search queries with other users.
- To eliminate the unwanted advertisement pop-ups.
- To improve the efficiency of the existing PWS by suggesting optimal solution.

4 Conclusion

The advancement in the technology like web search engine is boundless. Moreover the development of profile-based personalized search engine over a regular web search engine has catered many requirements of the user and inclusion of concept like privacy protection has served for the betterment of the system which has helped reduce various privacy concerns making the PWS more user-friendly. Though the existing PWS helps to retrieve relevant results to the user, there is still the need of improvising and providing more stable solution in order to make the whole system efficient and effective at a time. The other drawback of existing system is that it provides the privacy without knowing whether the user really want to personalize the information and other attributes like browsed query and visited links or share it with other registered users. Our proposed "Customized privacy" can help to overcome this drawback and let the user decide in case he/she wants to share the log with others and allows user to selectively share the information with related online services that is usually used for advertisement and research purposes. For future work, better generalizing strategies can be looked for that can replace the existing strategy and make the system more optimal and efficient.

References

1. Zhicheng Dou, Ruihua Song, Ji-Rong Wen: A Large-scale Evaluation and Analysis of Personalized Search Strategies. ACM Transactions 978-1-59593-654-7/07/0005 (2007)
2. Wilfred Ng, Lin Deng, Dik Lun Lee: Mining User Preference Using Spy Voting for Search Engine Personalization. ACM Transactions on Internet Technologies, Vol. 7, (2007)
3. Alexander Pretschner, Susan Gauch: Ontology Based Personalized Search. Proc. 11th IEEE Intl. Conf. on Tools with Artificial Intelligence, November (1999) 391–398
4. Andreas Krause, Eric Horvitz: A Utility-Theoretic Approach to Privacy in Online Services. Journal of Artificial Intelligence Research (2010) 633–662

5. Lidan Shou, He Bai, Ke Chen, Gang Chen: Supporting Privacy Protection in Personalized Web Search. IEEE Transactions on Knowledge and Data engineering, Vol. 26, NO. 2 (2014)
6. Xiaokui Xiao, Yufei Tao: Personalized Privacy Preservation. Proc. ACM SIGMOD, June (2006)

Part VII
Research on Optical Networks, Wireless Sensor Networks, VANETs, and MANETs

Traffic Classification Analysis Using OMNeT++

Deeraj Achunala, Mithileysh Sathiyanarayanan and Babangida Abubakar

Abstract There has been a lot of research on effective monitoring and management of the network traffic, where a large amount of internet traffic requires more accurate and efficient ways of traffic classification methods and approaches with an aim to improve network performance. In our research, we introduce the subject of packet classification in IP traffic analysis with a simple technique that relies on prototype classifier using OMNET++ (Optical Modelling Network using C++ programming language) which unfolds one new possibility for an online classification focusing on application detection in the absence of payload information. In this research, we evaluated our novel IATP (Inter-arrival time and precision) clustering algorithm with the help of OMNET++ scheduler for classification of network traffic. The analysis is based on the measure combined with inter-arrival time and precision which was able to distinguish fairly as a small different subset of clusters. With our implementation of a range of flow attributes, the simulation result demonstrates the effectiveness of 100% accuracy of classifying packets but does not constitute the same level of accuracy with real-time traffic classifier which operates under certain constraints. Accuracy for real-time traffic might normally varies from 80 to 95% and depends on the type of each application. Further study and heuristics are required for detecting much better methodologies for detecting applications with real-time traffic measurements.

Keywords Traffic classification · Clustering · OMNet++ · k-means · Quality of service (QoS)

D. Achunala (✉) · M. Sathiyanarayanan
School of Engineering, Swansea University, Swansea, UK
e-mail: deeraj.achunala@gmail.com

M. Sathiyanarayanan
e-mail: s.mithileysh@gmail.com

B. Abubakar
School of Computing, University of Brighton, Brighton, UK
e-mail: b.abubakar@brighton.ac.uk

P.K. Sa et al. (eds.), *Progress in Intelligent Computing Techniques: Theory, Practice, and Applications*, Advances in Intelligent Systems and Computing 719, DOI 10.1007/978-981-10-3376-6_45

1 Introduction

The internet complexity and scope is currently much faster than our ability to understand and predict it, especially with encrypted services such as video, P2P and VoIP. Due to the growing demand for bandwidth consumption and the introduction of new applications increase the importance of network traffic engineering, especially it is very useful to classify and analyse the network traffic independently for understanding spammers and malicious intruders. Therefore, an accurate classification and analysis of the network traffic flow is very much essential [1]. The earlier method for classifying the traffic was based on “payload information” and this method relied on some information of the payload format since every protocol decoding requires information of decoding the payload format and its characteristic patterns [2]. The main drawback of this approach is that the payload becomes inaccessible whenever there is an application implementation of a protocol change; it must be updated by an another classification procedure.

The classification concept we use in this work is to examine the unsupervised learning process since it is more advantageous to group data and has other practical benefits over labelled data. We are not using any real-time network traffic that includes any payload information. As an alternative we make use of machine learning methods based on the measurements of inter-arrival packet delays and packet lengths we expect to show a much better assumption and behaviour of different applications. Although the algorithm uses unsupervised learning mechanism, they are based on different principles of clustering like k-means, density-based spatial clustering of applications with noise (DBSCAN) and auto class algorithms [3, 4]. Usually, K-means and DBSCAN algorithms are chosen over auto class algorithm since they are capable of clustering the data much faster. The other classification method is a port-based classification. After classifying packet using this method, we have classification based on the flow information such as number of packets, mean inter-arrival time and duration, introduced by BLINC [2]. This research aims a fundamentally different approach to build a classifier to classify the internet traffic packets and then use clustering techniques based on different applications that generate them. We evaluate the performance parameters and the accuracy level of clustering the classified data. The simulation is carried out in OMNET++ tool and with the resulting data we plot the vectors in MATLAB. The purpose of this work is to produce a novel method that could bypass the limitations of the previous approaches.

In the next section of our paper, related work is explained to understand the deep packet inspection (DPI), the traditional methods of classification and the clustering algorithms. Based on the drawbacks of the previous methods, we came up with a novel algorithm called IATP (Inter-arrival time and precision) clustering algorithm with the help of OMNET++ scheduler for classification of network traffic. So, our traffic classification model is described in the later sections along with the performance evaluation. Finally, we conclude our discussions in this paper by identifying future works.

2 Related Work

Traffic classification is the first method which helps to identify different protocols and applications in a network. There has been a lot of research on effective monitoring and management of the network traffic, where a large amount of internet traffic requires more accurate and efficient ways of traffic classification methods and approaches with an aim to improve network performance. There are many generic classification methods and techniques described in [5].

Deep Packet Inspection. Deep packet inspection (DPI) is a form of packet filtering technique in computer networks which examines the information content or possibly the header of a packet as it traverses along the link. The inspection process decides whether a packet may be passed through or needs to be routed in a different path or to further inspect for the non-compliance of the protocol, intrusions, spam and viruses [6]. For traffic classification DPI is the foremost technology for authenticating protocols and identifying applications conveyed from an IP [1, 2]. DPI has been more troubling especially in traffic shaping and behavioural targeting (BT) [7].

Traditional methods of classification. The existing methods for traffic classification are based on three categories: (1) based on port numbers and the type of application, (2) based on the applicative layers, and (3) based on supervised learners—k-nearest neighbours (k-NN), discriminative analysis (DA) and support vector machines (SVM). Classification of traffic has played a vital role for numerous tasks such as QoS, trend analysis, dynamic access and lawful interception and monitoring. Traditionally, traffic classification was performed based on port and payload-based analysis but in the recent years machine learning techniques for classification have seen much higher interest. One such is a flow-based analysis.

Clustering Algorithms. In the clustering algorithms, we have K-means, density-based spatial clustering of applications with noise (DBSCAN) and autoclass. All these algorithms have some performance issues and drawbacks listed in [3, 4].

3 Traffic Classification Model

OMNeT++ [8] is an Objective Modular Network Testbed in C++, a discrete event simulator. OMNET++ modules can have parameters which are used mainly for three main purposes: to customize the behaviour of the model, for module communication as shared variables, and to create flexible model topologies (specifying the number of modules, connection structure, etc. by the parameters). Users are provided the lowest level of the module hierarchy containing the model of the algorithm. Simple modules appear to run in parallel, during simulation execution since they are implemented as co-routines. There is no need for a user to learn new programming language for writing simple modules, but the user need is expected to have a basic understanding of C++ programming. The simulator is basically portable since it is written in C++

and it should run on most platforms with a C++ compiler. The extended versions of OMNET++ are able to execute parallel simulations for any kind of synchronization. OMNET++ also provides explicit support for statistical synchronization.

Our main goal is to build an accurate and efficient classifier using clustering technique as depicted in Fig. 1a. To design such a model we consider two stages, model building as the first stage and classification as the second stage. In the model building stage the clustering algorithm clusters the training data and classification stage produces small subsets of clusters which are then labelled for our classification model. The simulation model can be used to cluster both online and offline models.

We combine the first- and second-stage techniques of traditional queuing (FIFO) First-In-First-Out discipline and our classifier design. The FIFO queues of finite size can have variable packets. In this model we use an automatic classification model where the classifier classifies the type of application flow into five different class IDs (Class 0 ID, Class 1 ID, Class 2 ID, Class 3 ID, and Class 4 ID). In OMNeT++, two integrated models are applied: the packet generator model and the packet scheduler and classifier model. Based on the type of the applications and protocol, the system classifies the packet flows into different classes, based on the Class ID (0–4).

4 Performance Evaluation

With all the results obtained from OMNET++ we export the data to MATLAB for evaluating the desired result of ‘clustering the packets as small subsets in our research. In the simulation process only the number of packets in one timeframe is considered for plotting the packets as inter-arrival time versus precision. The resulting plot is shown in Fig. 1b. Each resulting cluster can be differentiated as per the application implied. Clusters with its respective colours and packet counts are represented in the tables of Fig. 2. Using k-NN search option in MATLAB the distance between N packets from an X reference point in each cluster in ascending order is calculated as shown in the table of Fig. 2b. The analysis is based on the measure

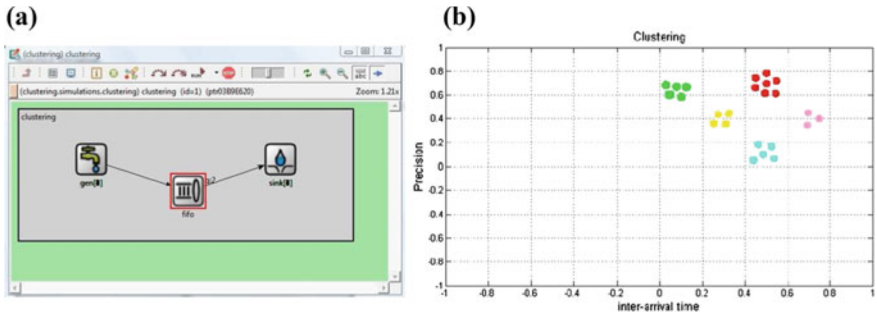
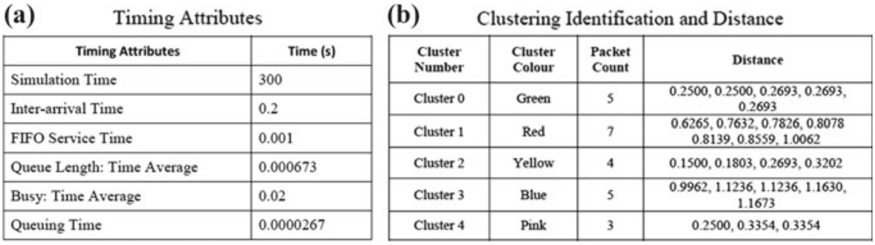


Fig. 1 **a** Traffic classification model designed in OMNet++. **b** Result of our classification



(a) Timing Attributes

Timing Attributes	Time (s)
Simulation Time	300
Inter-arrival Time	0.2
FIFO Service Time	0.001
Queue Length: Time Average	0.000673
Busy: Time Average	0.02
Queuing Time	0.0000267

(b) Clustering Identification and Distance

Cluster Number	Cluster Colour	Packet Count	Distance
Cluster 0	Green	5	0.2500, 0.2500, 0.2693, 0.2693, 0.2693
Cluster 1	Red	7	0.6265, 0.7632, 0.7826, 0.8078 0.8139, 0.8559, 1.0062
Cluster 2	Yellow	4	0.1500, 0.1803, 0.2693, 0.3202
Cluster 3	Blue	5	0.9962, 1.1236, 1.1236, 1.1630, 1.1673
Cluster 4	Pink	3	0.2500, 0.3354, 0.3354

Fig. 2 **a** Timing attributes. **b** Clustering identification and distance

combined with inter-arrival time and precision which was able to distinguish fairly as a small different subset of clusters. With our implementation of a range of flow attributes the simulation result demonstrates the effectiveness of 100% accuracy of classifying packets but does not constitute the same level of accuracy with real-time traffic classifier which operates under certain constraints.

5 Conclusion and Future Work

In our work we have evaluated an IATP (Inter-arrival time and precision) clustering algorithm with the help of OMNET++ scheduler for the classification of network traffic. The analysis is based on the measure combined with inter-arrival time and precision which was able to distinguish fairly as a small different subset of clusters. With our implementation of a range of flow attributes the simulation result demonstrates the effectiveness of 100% accuracy of classifying packets but does not constitute the same level of accuracy with real-time traffic classifier which operates under certain constraints. Accuracy for real-time traffic might normally vary from 80 to 95% and depends on the type of each application. Further study and heuristics are required for detecting much better methodologies for detecting applications with real-time traffic measurements. The project thesis has considered in-depth literature review [9].

Based on this survey on traffic classification we are still open to research on other issues as well on how to adapt and improve network services which we discuss in detail below. At first there is no defined level of measurement, and this has lead to other multi-dimensional problems with the existing equipments for measuring the traffic since the measurement levels are not thoroughly understood. As these classification techniques are still progressing to investigate with different methodologies. Our future work can include investigating of traffic classification with QoS since our models have been designed only with performance evaluation and not with QoS. In future we can evaluate QoS assurance with reliability, delay and throughput. By observing performance metrics such as classification rate and build time, a much better differentiation of algorithms can be investigated. Our future work will depend on the contributions and limitations of the other researchers work [10–15].

References

1. C. Parsons, *Deep Packet Inspection in Perspective: Tracing its lineage and surveillance potentials*. Citeseer, 2008.
2. T. Karagiannis, K. Papagiannaki, and M. Faloutsos, “Blinc: Multilevel traffic classification in the dark,” *SIGCOMM Comput. Commun. Rev.*, vol. 35, no. 4, pp. 229–240, Aug. 2005.
3. S. Valenti, D. Rossi, A. Dainotti, A. Pescapè, A. Finamore, and M. Mellia, “Reviewing traffic classification,” in *Data Traffic Monitoring and Analysis*. Springer, 2013, pp. 123–147.
4. J. Erman, M. Arlitt, and A. Mahanti, “Traffic classification using clustering algorithms,” in *Proceedings of the 2006 SIGCOMM Workshop on Mining Network Data*, ser. MineNet ’06. ACM, 2006, pp. 281–286.
5. A. W. Moore and D. Zuev, “Internet traffic classification using bayesian analysis techniques,” in *Proceedings of the 2005 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*, ser. SIGMETRICS ’05. ACM, 2005, pp. 50–60.
6. P. Piyachon and Y. Luo, “Efficient memory utilization on network processors for deep packet inspection,” in *Proceedings of the 2006 ACM/IEEE symposium on Architecture for networking and communications systems*. ACM, 2006, pp. 71–80.
7. A. Bremler-Barr, Y. Harchol, D. Hay, and Y. Koral, “Deep packet inspection as a service,” ser. CoNEXT ’14. ACM, 2014, pp. 271–282.
8. A. Varga. OMNeT++: Discrete event simulation system. [Online]. Available: http://www.omnetpp.org/.
9. D. Achunala, “Traffic classification,” Sep 2012.
10. A. Bremler-Barr, S. T. David, Y. Harchol, and D. Hay, “Leveraging traffic repetitions for high-speed deep packet inspection,” in *Computer Communications (INFOCOM), 2015 IEEE Conference on*. IEEE, 2015, pp. 2578–2586.
11. J. Sherry, C. Lan, R. A. Popa, and S. Ratnasamy, “Blindbox: Deep packet inspection over encrypted traffic,” in *Proceedings of the 2015 ACM Conference on Special Interest Group on Data Communication*. ACM, 2015, pp. 213–226.
12. C. Hu, H. Li, Y. Jiang, Y. Cheng, and P. Heegaard, “Deep semantics inspection over big network data at wire speed,” *Network, IEEE*, vol. 30, no. 1, pp. 18–23, 2016.
13. M. Sathiyanarayanan and K. S. Kim, “Multi-channel deficit round-robin scheduling for hybrid tdm/wdm optical networks,” in *Proc. of the 4th International Congress on Ultra Modern Telecommunications and Control Systems (ICUMT 2012)*, St. Petersburg, Russia, Oct. 2012, pp. 552–557.
14. M. Sathiyanarayanan and B. Abubhakar, “Dual mcdrr scheduler for hybrid tdm/wdm optical networks,” in *Proc. of the 1st International Conference on Networks and Soft Computing (ICNSC 2014)*, Andra Pradesh, India, Aug 2014, pp. 466–470.
15. M. Sathiyanarayanan and B. Abubakar, “Mcdrr packet scheduling algorithm for multi-channel wireless networks,” in *Proceedings of 3rd International Conference on Advanced Computing, Networking and Informatics*. Springer, 2016, pp. 125–131.

Irregular-Shaped Event Boundary Estimation in Wireless Sensor Networks

Srabani Kundu, Nabanita Das, Sasanka Roy and Dibakar Saha

Abstract In a wireless sensor network (WSN), sensor nodes are deployed to monitor a region. When an event occurs, it is important to detect and estimate the boundary of the affected area and to gather the information to the sink node in real time. In case, all the affected nodes are allowed to send data, congestion may occur, increasing path delay, and also exhausting the energy of the nodes in forwarding a large number of packets. Hence, it is a challenging problem to select a subset of affected nodes, and allow them only to forward their data to define the event region boundary satisfying the precision requirement of the application. Given a random uniform node distribution over a 2-D region, in this paper, three simple localized methods, based on local convex hull, minimum enclosing rectangle, and the angle of arrival of signal, respectively, have been proposed to estimate the event boundary. Simulation studies show that the angular method performs significantly better in terms of area estimation accuracy and number of nodes reported, even for sparse networks.

Keywords WSN · Event area · Convex hull · Boundary detection · Minimum enclosing rectangle

S. Kundu (✉)
Guru Nanak Institute of Technology, Kolkata, India
e-mail: srabani6@gmail.com

S. Kundu · N. Das · S. Roy · D. Saha
Indian Statistical Institute, Kolkata, India
e-mail: ndas@isical.ac.in

S. Roy
e-mail: sasanka.ro@gmail.com

D. Saha
e-mail: dibakar.saha10@gmail.com

P.K. Sa et al. (eds.), *Progress in Intelligent Computing Techniques: Theory, Practice, and Applications*, Advances in Intelligent Systems and Computing 719, DOI 10.1007/978-981-10-3376-6_46

1 Introduction

In general, to monitor large inaccessible regions, wireless sensor networks are deployed with tiny, inexpensive sensor nodes distributed over an area to collect ground data [1, 2]. At regular intervals, the nodes sense data and forward it to the sink node via multihop paths.

A sensor node is basically a small device capable of sensing data, some processing, and communicating with its neighboring nodes. Here, the sensor nodes are assumed to be homogeneous and static. In most of the cases nodes are battery powered with limited or no recharging facility at all. Also, the computing capability of a node is elementary with small amount of storage. It is to be noted that typically, communication demands most of the energies of a node, whereas sensing and computing take only a small share. So, to enhance the network lifetime, it is extremely essential to limit the number of packets in the network. This necessitates in-node processing, i.e., instead of forwarding the incoming packets to the sink continuously, nodes may process data and forward the relevant information only toward the sink node. However, with limited computing power and limited memory, the in-node processing should be simple in terms of computation complexity and storage.

When an event occurs within the area to be monitored as shown in Fig. 1, it may spread over the region and it should be identified immediately. If all affected nodes start to route their information to the sink node, the network gets congested immediately resulting increase in packet delay. Also, due to huge number of packet forwarding, nodes will die out faster which can create a network failure. Hence, it is always better to choose a small subset of affected nodes which are critical to reconstruct the event region boundary, and to allow them to send their packets to the sink node only. The reduction in the number of reporting nodes at one hand limits the traffic in the network, saving energy significantly. On the other hand it also helps to reduce

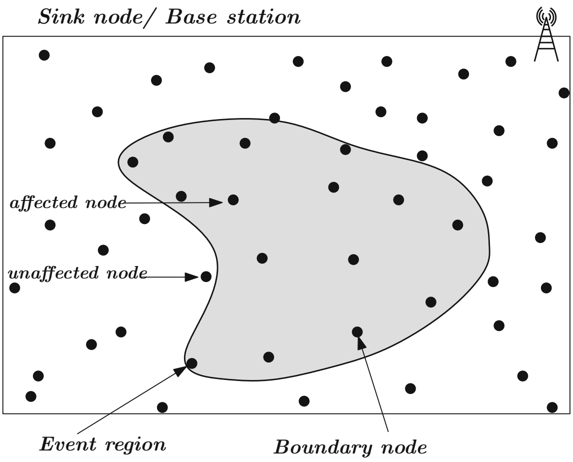


Fig. 1 Event boundary and affected nodes

congestion, hence path delay for real-time reporting. Again, since it results some loss of information, it is challenging to optimize the number of reporting nodes to satisfy the precision requirement of the concerned application. Knowing all the affected nodes, the problem can be easily mapped to the classical problem of computational geometry, namely the convex hull computation. However, it is to be noted that in WSNs instead of optimal centralized algorithms, it is wise to adopt self-organized light-weight localized algorithms based on local neighborhood information only that converges with limited rounds of communication.

A lot of research activities have been reported so far on event boundary estimation problem in WSN, formulated in various ways to combat their inherent hardness. The important challenges are to limit the amount of computation and rounds of communication, and at the same time the computation should be based on minimum neighborhood information, since message communication is the only way to gather knowledge about the neighborhood of a node, and it is expensive in terms of energy. Authors in [3] presented a boundary estimation method based on two centrality measures of nodes, betweenness and closeness, respectively. In [4, 5], a graph-theory-based solution is developed to detect the event boundary, irrespective of any communication model. Based on the concept of image processing, Chintalapudi et al. in [6] proposed an algorithm to detect the network boundary. Another statistical approach to identify the boundary nodes and the topology of the region has been presented in [7].

Authors in [8, 9] proposed techniques based on computational geometry. A polynomial-based boundary estimation algorithm has been proposed in [10], where the *query tree* was constructed to route the event information in the form of a polynomial to the *sink* node. In [11–13], authors proposed some heuristic-based solutions to detect and identify the event boundary for a wireless sensor network. The gradient-based data distribution model is followed by the authors in [14, 15] to detect the event boundary for an irregular-shaped event area. In [16, 17], authors proposed a low latency event boundary detection heuristic where it generates a reduced boundary node set, without knowing the neighbors' locations and forward it to the sink node with minimum latency. Most of the above algorithms are either computation intensive, or are based on many unrealistic assumptions. In WSN, the sensed data are highly error prone and the assumption of graded data distribution is not always true. Again, detection of the boundary with the help of neighbor nodes location information requires large memory which is really very difficult to manage.

Considering a uniform random node distribution over a 2-D region, in this paper, we focus on three simple distributed methods to estimate the irregular-shaped event boundary region in WSN. First, we present two naive techniques—one based on localized convex hull, and the minimum enclosing rectangle, respectively. Finally, we propose a simple light-weight distributed algorithm based on angle of arrival of signal with $O(d \log d)$ computation and $O(d)$ space complexity in each affected node, where d is the maximum number of neighbors of a node. Each node is assumed to be equipped with directional antenna and is capable of measuring the angle of arrival of received signal. For in-node processing, each node requires the node ids of its adjacent neighbors, and only their locations are not required. Extensive simulation studies show that the angular boundary detection method needs minimum computation and communication overhead and it also can detect the event boundary more accurately compared to the others even when the region is sparsely populated. The paper is organized as follows. Section 2 defines the problem. Section 3 proposes the distributed algorithms for the selection of the boundary nodes. Section 4 shows the simulation results and finally, Sect. 5 concludes with some open issues.

2 Network Model and Preliminaries

In our model of wireless sensor networks, the 2-D region under consideration is deployed with n homogeneous sensor nodes, randomly distributed over the area. Each node i can communicate directly with a node j if it lies within its transmission range T.

Definition 1 Two sensor nodes i and j are *neighbors* of each other, if and only if, sensor node i can communicate with node j directly, i.e., the Euclidean distance $D(i,j)$ between nodes i and j is less than the *transmission range* T, i.e., $D(i,j) \leq T$.

Definition 2 A WSN is represented by an undirected *topology graph* $G(V,E)$, where V is the set of nodes distributed over a 2D region and E is the set of edges such that an edge $(i,j) \in E$, if and only if j is a *neighbor* of i and vice versa, with $i,j \in V$.

Definition 3 In a *topology graph* $G(V,E)$, the *hop count* of a node i is represented as its distance in terms of number of hops from the sink node via shortest path.

Let us assume that each sensor node senses the environment at a regular interval of time and when required, routes the sensed data to the sink node. When an event occurs, in general, it spans over a region which may be of irregular shape. To estimate the event region boundary of irregular shape, by selecting a few boundary nodes only, it is an important and challenging problem in WSN. With the above network model, we consider the problem of selecting a reduced set of boundary nodes in a distributed fashion, which reports to the sink node which reconstructs the event boundary and estimates the affected area in terms of the convex hull enclosing all reported nodes.

Definition 4 Given a set of points S distributed over a $2-D$ region, the convex hull of S is defined as the smallest convex polygon enclosing all points of S.

To achieve the solution with acceptable accuracy level, we propose three distributed algorithms with simple in-node processing based on limited neighborhood information that converges with small number of communication rounds. In our proposed model, sensor nodes do not require the actual data value, and no assumption has been taken about the data distribution within the event area.

3 Algorithms for Estimating the Irregular-Shaped Event Boundary Region

To detect the change of environmental phenomenon at a regular interval of time, we assume that sensor nodes are deployed randomly over an area. When an event occurs, it spans an area R of arbitrary shape without any hole. If the sensed data crosses a threshold value, then a node executes the boundary detection algorithm as described below. We propose three simple distributed schemes for selecting the boundary nodes and compare their performance by simulation.

3.1 Boundary Detection by Localized Convex Hull

Here, we present a distributed algorithm based on localized convex hull computation to detect the boundary nodes of an event region $\mathcal{R}$ as shown in Fig. 2. Here, each affected node i detects all its affected neighbors and constructs a local convex hull by considering all its affected neighbors with their locations. If node i itself is one of the vertices of the convex hull, node i announces itself as a boundary node, and forwards its location to the sink. For routing, a spanning tree may be constructed in the WSN to forward data via minimum delay path as has been proposed in [17]. Initially, each node broadcasts a 'Hello' packet with its node id and location, and from the 'Hello' packets received from others it prepares neighbor list with their locations. Each node senses data at regular interval; in case it exceeds the predetermined threshold value, it broadcasts an 'Affected' message with its node id and location. From the

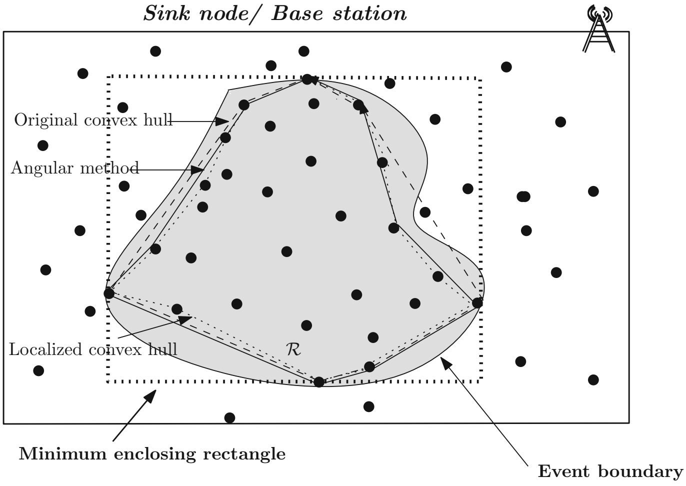


Fig. 2 Boundary detection by convex hull, minimum enclosing rectangle, and angular method

received ‘Affected’ messages from neighbors, it computes the convex hull enclosing itself and all its affected neighbors by the well-known *Jarvis March* algorithm [18]. The algorithm is the simplest one for constructing convex hull enclosing points on a two-dimensional plane with $O(h.n)$ time complexity, where h is the number of vertices of the convex hull and n is the number of points given. In real-life examples, the *Jarvis March* algorithm outperforms other convex hull algorithms when n is small or h is expected to be very small compared to n. In our case, n is limited by the maximum node degree d, and $h \leq d$, hence in the worst case, the complexity is $O(d^2)$. The space complexity of the procedure is $O(d)$ only. It is evident that with a collision-free message protocol, each node transmits only 3 messages, and the procedure terminates in 3 rounds only.

Therefore, the above procedure is simple, with $O(d^2)$ time complexity, and constant message complexity. Each node takes the decision of selection by itself based on the locations of its affected neighbors only. Also, the procedure converges in 3 rounds only. But the performance in terms of accuracy in boundary estimation is not guaranteed.

3.2 Boundary Detection by Minimum Enclosing Rectangle

By the most naive approach, the event area is estimated by finding the minimum rectangle enclosing all affected nodes. For this, the sink node should know the extreme co-ordinates of the affected nodes. Each affected node i knows its co-ordinates (x_i, y_i), and sets $x_{min}(i) = x_{max}(i) = x_i$ and $y_{min}(i) = y_{max}(i) = y_i$, and broadcasts it. Next, it listens to its neighbors. Each time if it receives a packet from its neighbor j and if $x_{min}(j) < x_{min}(i)$, then $x_{min}(i) \leftarrow x_{min}(j)$ and it is broadcasted. If $y_{min}(j) < y_{min}(i)$, then $y_{min}(i) \leftarrow y_{min}(j)$, and then it is broadcasted. Similarly, if $x_{max}(j) > x_{max}(i)$, then $x_{max}(i) \leftarrow x_{max}(j)$. If $y_{max}(j) > y_{max}(i)$, then $y_{max}(i) \leftarrow y_{max}(j)$. If there is any update it is broadcasted. If any unaffected node receives any updated value of the four variables mentioned above, it broadcasts it. The procedure terminates after P rounds of communication, where P is the maximum hop count of a node in $G(V, E)$. Finally, the sink computes the minimum enclosing rectangle with x_{min}, y_{min}, x_{max} and y_{max}. Figure 2 shows an event area enclosed by the minimum enclosing rectangle. Though the computation involved is very simple, but gathering of the extreme co-ordinates of the affected nodes necessarily requires flooding in the network that in the worst case may take P rounds of communication. The computational complexity of each node is $O(P.d)$. The message complexity is $O(P)$. It needs only the information of its own location. It is clear that this approach always over estimates the area, and the convergence is rather slow. In the worst case it may take $O(n)$ rounds to complete.

3.3 Angular Boundary Detection

Finally, we propose another approach, based on the angular location of the neighbors. It is assumed that each node is equipped with directional antenna, such that when it receives a signal, it can estimate the angle of arrival. Hence, each affected node selects a subset of its affected neighbors as the reporting nodes. By *Angular boundary detection algorithm*, each affected sensor node i broadcasts a *Hello* $(i, \mathit{flag} = 1)$ message with its id to its neighbors and listens from its neighbors ($\mathit{flag} = 0$ means unaffected node). If it receives a *Hello* $(j, \mathit{flag} = 1)$ message from its neighbor j, it just includes it in its neighbor list *NL* with its id, flag bit, and the angle of arrival θ_j.

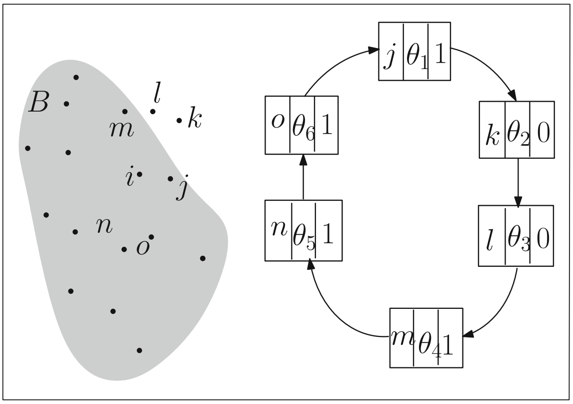
Next, each affected node i checks whether all of its neighbors are affected or not. If not, node i includes its neighbors in a circular list L in sorted order of θ_j (may be clockwise or anticlockwise) as shown in Fig. 3. Finally, node i starts to traverse L and checks for any transition from affected node to unaffected node or vice versa. If any transition is found then the affected node $j \in L$ is selected as boundary node, and it is added to a list and finally node-i broadcasts the list. Each node, if selected, forwards its location to the sink.

Algorithm 1 presents the steps formally.

Complexity Analysis:

- **Time complexity**: Each node computes the neighbor positions in terms of angles and sort them in a list L. After sorting, each node traverses the list only once. Assuming that in $G(V, E)$ the maximum node degree is d, each node, in the worst case, requires $O(d \log d)$ computation.
- **Space complexity:** Each node makes a list of neighbors *NL*. To make the list L, each entry consists three elements, i.e., node id, flag, and angle. If all the d neighbors get included in the list, we need $O(d)$ storage space. Hence, the space complexity is $O(d)$.

Fig. 3 Affected node i and its neighbors in L



Algorithm 1: Angular Boundary Detection

```
Input: Node $i$, $STATUS = 0$, $flag = 0$, list of neighbors $NL$
Output: $STATUS = 0/1$ of a node (boundary node or not)
for each node i do
    if an affected node then
        $flag \leftarrow 1$;
    end
    phase 1: node $i$ broadcasts a $hello(i, flag = 1)$ message ;
    wait and listen;
    if receives a $hello(j, flag = 1/0)$ from its neighbor $j$ then
        include $j$ in $NL$ with its $id$, $flag$ bit and angle of arrival;
    end
    Phase 2:
    for each node-$j \in NL$ do
        if $flag == 1$ then
            $temp \leftarrow 1$;
        else
            $temp \leftarrow 0$;
            break;
        end
    end
    if $temp == 0$ then
        for each node $j \in NL$ do
            The angle of arrival $\theta$ is included in $L$ in sorted order;
        end
    end
    Phase 3: $temp \leftarrow flag$// the flag value is the first node $j \in L$;
    for each node-$j \in L$ do
        if $temp != flag$ //transition found then
            include affected node $j$ in a temporary list $L_t$ // for boundary node;
            $temp \leftarrow flag$;
        end
    end
    broadcasts selected($L_t$) message;
    if receives selected($L_t$) message then
        if $STATUS = 0$ then
            if $i \in L_t$ and then
                $STATUS \leftarrow 1$;
            end
        end
    end
    Terminate;
end
```

- **Message complexity**: Only two messages per node are required in the procedure, one *Hello* message and one *selected* message. Hence, per node message complexity is $O(1)$.

Example 1 Figure 3 shows an arbitrary event boundary B. In this example, node i collects angular location information from its 6 neighbor nodes and constructs a circular list by sorting the angles in anticlockwise direction as shown in Fig. 3. Now, node i starts to traverse through the list and finds transitions from node j to k and from l to m. As node j is affected, so node i declares node j as the boundary node. Similarly node i also declares node m as another boundary node.

4 Simulation Studies

For simulation study, given a $w \times w$ square area A with a random uniform distribution of n nodes, irregular-shaped event area is generated by diffusion process model following [15].

4.1 Arbitrary Event Area Generation

In [15], the event area is generated by two steps, *diffusion* and *softening*. Here, the entire area to be monitored is divided into $w \times w$ grid. Next, some grid cells are randomly chosen as *source cells* and initialized with a high data value. The cells other than the *source cells* are initialized to a fixed lower data value. In diffusion step, keeping the data of the *source cells* unaltered, the data values of all other cells are updated by the average of its four neighbor cells. After repeated application of this diffusion step, softening step is followed where the sources became non-source cells and some cells are again randomly chosen as sources except those previous cells. The process is repeated to generate the event area. In this work, we have customized this procedure to generate our event area within 200×200 grid. Here source cells are chosen randomly and they are adjacent with each other. Then the diffusion and softening steps are being carried out to generate the event area as shown in Fig. 4.

4.2 Results

For simulation, $1000 \leq n \leq 2500$ homogeneous nodes are distributed over the 200×200 region by considering a uniform random distribution. Here, different event

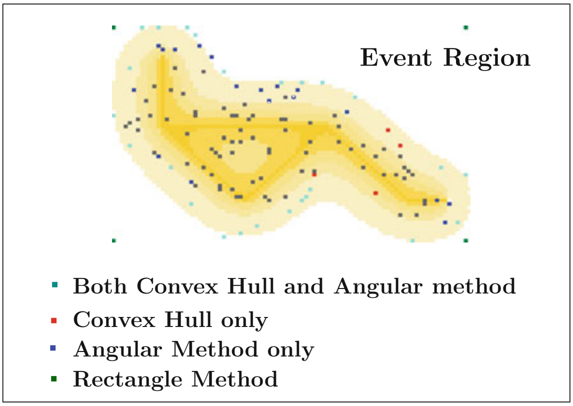


Fig. 4 Event area generated by diffusion model

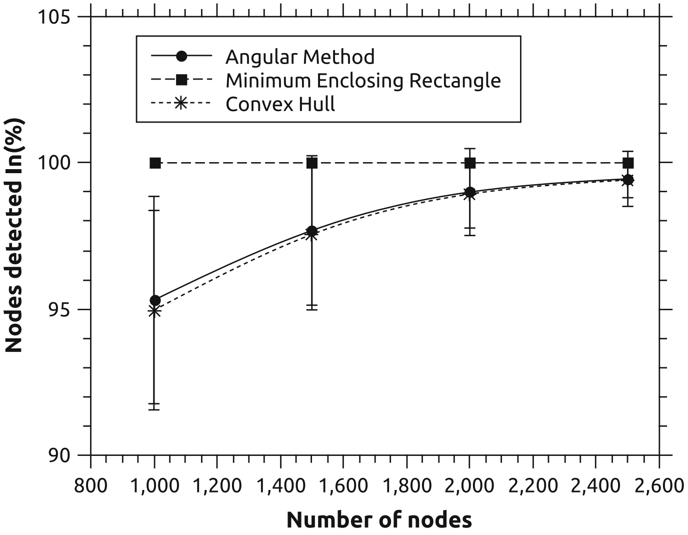


Fig. 5 Affected boundary nodes enclosed (%) versus *n*

regions are created by changing the source cells randomly and the experiments are repeated for different networks by varying the node set and the transmission radius. The transmission radius T varies between $6 \leq T \leq 12$. The simulation is implemented using Java 1.7.0_55 and Matlab.

Figure 5 shows how the percentage of affected nodes enclosed within the estimated area varies with n. It is evident that minimum enclosing rectangle method always encloses 100 % of affected nodes, whereas, for the other two methods, the percentage increases with n as is expected. On the other hand, Fig. 6 shows the variation of the number of unaffected nodes enclosed within the estimated boundary, termed here as *false positive*. For the rectangle method, the false detection rate is very high which will always over estimate the event area.

From the simulation studies, it is also clear that the angular method performs well even with low node density which is very suitable for real-life scenario.

Figure 7 shows that the angular boundary detection method reports small number of boundary nodes compared to the convex hull procedure, in case the node density is low. It is also evident from Fig. 8 that with small number of boundary nodes angular boundary detection method always gives better accuracy of area estimation compared to the other two methods.

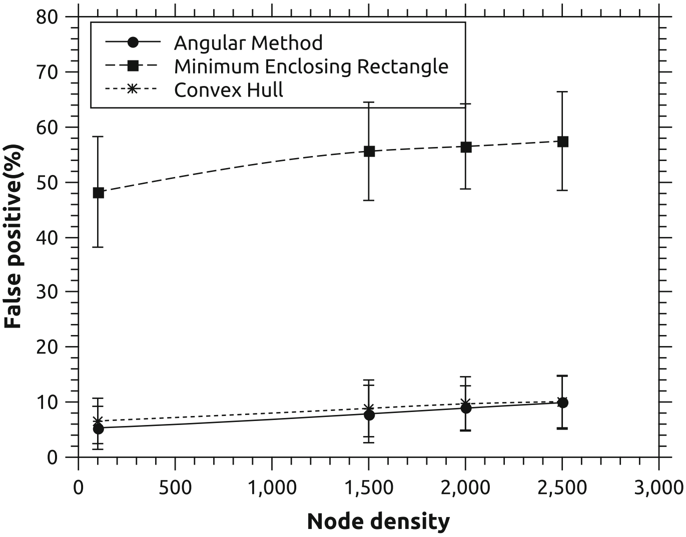


Fig. 6 Unaffected nodes enclosed (in %) versus *n*

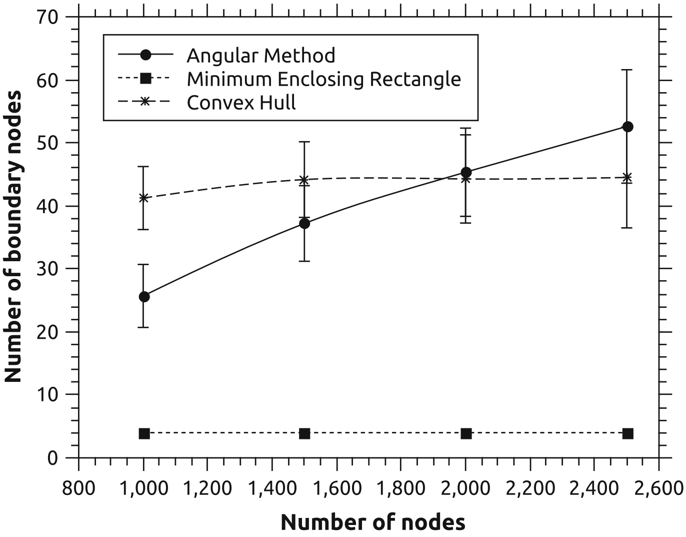


Fig. 7 Boundary nodes reported (in %) versus *n*

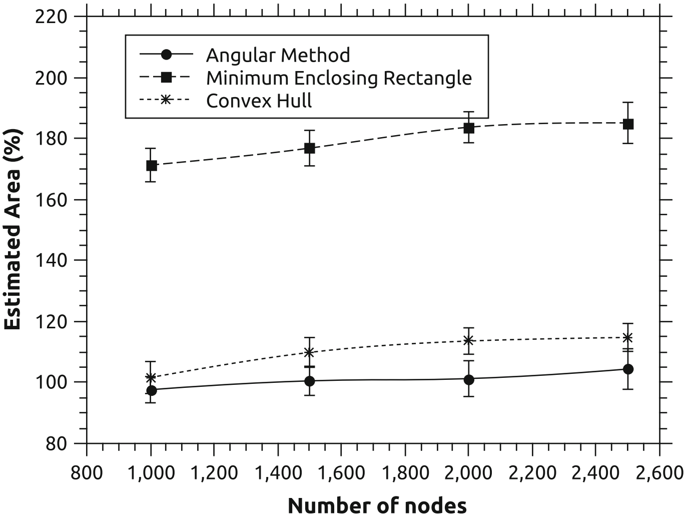


Fig. 8 Estimated area (in % of actual area) versus *n*

5 Conclusion and Future Work

Given a random node distribution over a bounded 2-D area, we focus on simple distributed approaches to estimate the irregular-shaped event boundary region in wireless sensor networks. We propose three algorithms, namely the (a) localized convex hull, (b) the minimum enclosing rectangle, and (c) the angular boundary detection. Complexity analysis (both time and message) and comparison studies by simulation show that the proposed angular boundary algorithm, without neighborhood location information, performs better in terms of accuracy in event boundary detection, number of reported boundary nodes, and rounds of communication.

References

1. A. Mainwaring, J. Polastre, R. Szewczyk, D. Culler, and J. Anderson, *Wireless sensor networks for habitat monitoring*, In Proceedings of the 1st ACM International Workshop on Wireless Sensor Networks and Applications (Atlanta, Sept.). ACM Press, New York, 2002.
2. I.F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, *Wireless sensor networks: a survey*, Computer Networks Journal, Elsevier, March, 2002.
3. X Li, S He, J Chen, X Liang, R Lu and S Shen, *Coordinate-free Distributed Algorithm for Boundary Detection in Wireless Sensor Networks*, In Proceedings. of IEEE Globecom Houston, TX, USA Dec. 2011.
4. Y. Wang, J. Gao and J. S. B. Mitchell, *Boundary Recognition in Sensor Networks by Topological Methods*, In MobiCom'06 Los Angeles, California, USA Sep. 2006.
5. R. Ghrist and A. Muhammad, *Coverage and Hole Detection in Sensor Networks via Homology*, In 4th international Symposium on Information Processing in Sensor Networks, April 2005.

6. K. K. Chintalapudi and R. Govindan, *Localized edge detection in sensor fields*, In Proc. Ad Hoc Networks, Elsevier, 2003.
7. S. P. Fekete, A. Kroller, D. P. Fisterer, S. Fischer and C. Buschmann, *Neighborhood-Based Topology Recognition in Sensor Networks*, In Proceedings ALGOSENSORS, Springer LNCS 2004.
8. Q. Fang, J. Gao and L. Guibas, *Locating and Bypassing Routing Holes in Sensor Networks*, In Proceedings Mobile Networks and Application, Springer, 2006.
9. C. Zhang, Y. Zhang, Y. Fang, *Localized algorithms for coverage boundary detection in wireless sensor networks*, In Wireless Network, Springer, 2007.
10. T Banerjee, D Wang, B Xie and D. P. Agarwal, *PERD: Polynomial Based Event Region Detection in Wireless Sensor Network*, In Proceedings of IEEE International Conference on Communications 2007.
11. J. Bruck, et al., *Localization and routing in sensor networks by local angle information*, ACM Transaction on Sensor Networks, 2009.
12. W. C. Chu, K. F. Ssu, *Decentralized Boundary Detection without Location Information in Wireless Sensor Networks*, In Proceedings IEEE Wireless Communications and Networking Conference: Mobile and Wireless Networks, 2012.
13. B. Greenstein, E. Kohler, D. Culler, and D. Estrin, *Distributed Techniques for Area Computation in Sensor Networks*, In Proceedings of the 29th Annual IEEE International Conference on Local Computer Networks (LCN'04).
14. S. Kundu and N. Das, *In-Network Estimation and Localization in Wireless Sensor Networks*, In Proceedings 7th IEEE International Conference, Globecom, Dec. 2012.
15. J. Lian, L. Chen, K. Naik, Y. Liu and G. B. Agnew, *Gradient Boundary Detection for Time Series Snapshot Construction in Sensor Networks*, IEEE Transaction on Parallel and Distributed Systems Oct. 2007.
16. S. Kundu and N. Das, *Event Boundary Detection and Gathering in Wireless Sensor Networks*, In Proceedings 2nd International Conference Applications and Innovations in Mobile Computing (AIMoC 2015), Feb. 2015.
17. S. Kundu, *Low Latency Event Boundary Detection in Wireless Sensor Networks*, In Proceedings International Conference on Advanced Networks and Telecommunications System (ANTS), 2015.
18. R. A. Jarvis, *On the identification of the convex hull of a finite set of points in the plane*, Information Processing Letters 2: 18–21, (1973). doi:10.1016/0020-0190(73)90020-3.

A New Two-Dimensional Mesh Topology with Optical Interlinks

Amritanjali

Abstract The performance of parallel computer heavily depends on the topology of the interconnection network. Two-dimensional mesh is a well-known topology for processor arrays. However, its large diameter increases execution time when the parallel algorithm requires communication between arbitrary pair of nodes. Wraparound connections between end nodes reduces its diameter, however, increases the complexity in the design of parallel algorithms. In this paper, we have proposed an intermediate approach, where additional links are used to reduce the diameter without increasing the design complexity. These additional optical links provides high-speed communication between nodes that are separated by half the number of nodes in each dimension. Also, we present efficient parallel algorithms for some elementary problems on the proposed system.

Keywords Parallel algorithms · Interconnection networks · Mesh topology

1 Introduction

The execution time of a parallel program not only depends on the number of computational steps that each processor has to execute but also on the time spent in communication as all the processors are working together to achieve a common goal. Communication between parallel processes takes place through shared memory or by sending messages to each other. In SIMD machines, like processor arrays, the data is distributed among the local memories of different processors. The processors can communicate with each other by sending messages, using the

Amritanjali (✉)
Department of Computer Science and Engineering, Birla Institute of Technology,
Mesra, Ranchi, India
e-mail: amritanjali@bitmesra.ac.in

P.K. Sa et al. (eds.), *Progress in Intelligent Computing Techniques: Theory, Practice, and Applications*, Advances in Intelligent Systems
and Computing 719, DOI 10.1007/978-981-10-3376-6_47

interconnection network. Only the processors that are neighbors can directly communicate. MIMD machines can be distributed memory or shared memory systems. Nowadays, the hybrid model has become more popular and the fastest computers of today use both shared and distributed memory architectures.

The interconnection network used in the parallel computer plays a key role in determining the overhead incurred on running a parallel algorithm on it. The topology of an interconnection network is described using graph, where nodes are the processing elements and the edges are the links between them. Mesh, tree, hypercube, butterfly, and shuffle-exchange networks are some of the well-known static topologies. Every topology has some advantage and disadvantage. No topology gives optimal performance under all conditions.

Mesh topology in particular is quite popular and several of the commercial available parallel computers [1–4] are based on it because of its regularity and low hardware complexity. In the basic mesh, topology nodes are organized in the form of q-dimensional lattice. The nodes are connected by links which can be unidirectional or bidirectional. Additionally wraparound links can be used to connect the border nodes at the opposite end. The mesh network with wraparound connections is called as torus. The diameter of a q-dimensional (without wraparound connections) with k nodes in each dimension is given by q(k − 1), as we have to travel at least (k − 1) edges to reach from one corner to another corner in each dimension. Bisection width is high, k^{q-1}. Node degree is 2q and for two and three-dimensional mesh maximum edge length is constantly independent of network size. The large value of diameter is the main drawback of this topology. Various mesh-based hybrid networks were introduced for better performance, like Mesh of Trees, Multi–Mesh, and Multi-mesh of Trees [5–7]. Recently, optical links have become quite popular as an interconnection medium in parallel systems for performing high-speed computing [8]. OTIS-MOT [9] and OMULT [10] are some the hybrid topologies using optical links. Some of the advantages of optical links are increased communication speed, reduced power consumption, etc. We have used the optical links in the basic mesh topology to improve communication efficiency. The proposed mesh interconnection network and its properties are described in Sect. 2 of the paper. Section 3 presents some elementary algorithms designed for proposed topology. Finally, we conclude our work in Sect. 4.

2 Two-Dimensional Mesh with Optical Interlinks

The nodes in the proposed topology are organized in the form of n x n two-dimensional array, where n is power of 2. The processor are numbered row-wise such that, P_{1n} is at the top right corner, P_{n1} at the bottom left corner and P_{nn} at the bottom right corner. In addition to the normal electronic links between adjacent processors, optical links are used to provide more connectivity and reduce

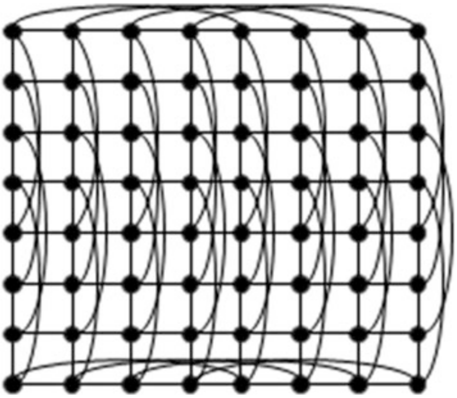


Fig. 1 8 × 8 Mesh with optical interlinks. Horizontal optical links are shown only for first and last row for clarity

diameter. Also, the optical links are faster to the electronic links. In each row i, where 1 ≤ i ≤ n, there are horizontal optical links between processor P_{ij} and $P_{i\ (j+n/2)}$, where 1 ≤ j ≤ n/2. For each column j, where 1 ≤ j ≤ n, there are vertical optical links between processor P_{ij} and $P_{(i+n/2)j}$, where 1 ≤ i ≤ n/2. Figure 1 shows the topology of an 8 × 8 mesh with optical interlinks.

Diameter. In a two-dimensional mesh with n^2 nodes, the distance between two corner nodes in the same row/column is (n/2 − 1) electronic links and 1 optical link, i.e., n/2 links. Therefore, diameter of the two-dimensional n x n mesh is n.

Bisection Width. To divide the network into two equal parts, we need to remove n/2 optical links in each row/column, in addition to n electronic links. Hence, bisection width is $n^2/2 + n$.

Node Degree. Maximum number of links per node is 6, 4 electronic links and 2 optical links.

Maximum Edge Length. Length of the optical link increases with increase in the size of the network. It connects nodes that are n/2 distance apart in n x n mesh.

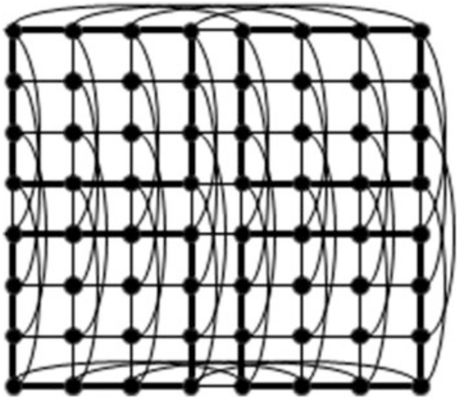
3 Some Elementary Parallel Algorithms

We present implementation of parallel algorithms for some elementary problems on the proposed mesh topology.

3.1 Data Broadcasting

The nodes in the n × n network can be divided into four groups of n/2 × n/2 nodes as shown in Fig. 2. To broadcast a data from any node of a group to all the nodes in the network, it is first broadcasted to all the nodes in its group using

Fig. 2 Dividing the network into 4 groups of processors, the border edges of each group are shown in bold



electronic links in maximum of 2(n/2 − 1) steps. Then, using horizontal optical links the data is sent to the nodes in the group adjacent to it horizontally. Finally, using the vertical optical links it is sent to the nodes in the remaining two groups. Therefore, one to all broadcasting can be done in n times.

3.2 Summation

We have n^2 data elements stored in the nodes of the mesh network. In the first phase of parallel summation, the sum is done in each group, row-wise, bringing the partial sum to the first column of each group, using the electronic links in (n/2 − 1) communication steps. Next, the horizontal optical links are used to bring the new partial sums in the first column of the mesh. Then, using the vertical optical the partial sums are brought to the first column of the first group. In the last phase, the final sum is produced at the processor P(1, 1) using the (n/2 − 1) electronic links. Hence, the parallel sum is generated in n communication steps.

Similarly, we can find average, maximum or minimum operations of n^2 elements in n communication steps.

3.3 Prefix Computation

The elements are distributed over the n^2 processors in the network, in row major order. First the prefix sum computation is done on each row in parallel, and then the results are combined. We assume that all the processors have three registers, tmp, p_sum and prefix.

Algorithm:

```
 {prefix computation for each row}
  for k ← 1 to n/2 -1 do
     for all $P_{ij}$, where $1 \le i, j \le n$, do
        if (j > k) then
         tmp ⇐ [i, (j–1)]prefix
         prefix ← prefix + tmp
       endif
    endfor
 endfor

 {using horizontal optical links}
 for all $P_{ij}$, where $1 \le i \le n$ and $(n/2+1) \le j \le n$ do
       tmp ⇐ [i, (j-n/2)]prefix
       prefix ← prefix + tmp
 endfor

 {prefix computation on the last column of every row}
 for all $P_{in}$, where $1 \le i \le n$, do
      p_sum ← prefix
 end for
 for k ← 1 to n/2 -1 do
    for all $P_{in}$, where $1 \le i \le n$, do
       if (i > k) then
        tmp ⇐ [(i-1), n]p_sum
        p_sum ← p_sum + tmp
       endif
    endfor
 endfor

 {using vertical optical links}
 for all $P_{in}$, where $(n/2 + 1) \le i \le n$ do
       tmp ⇐ [(i - n/2), n]prefix
       prefix ← prefix + tmp
 endfor


 {add p_sum to prefix of all the nodes in the same row}

 for all $P_{in}$, where $1 \le i \le n$, do
      prefix ←  prefix + p_sum
 end for

 for k ← (n-1) down to (n/2+1) do
    for all $P_{ik}$, where $2 \le i \le n$, do
       p_sum ⇐ [i, (k+1)]p_sum
       prefix ← prefix + p_sum
    endfor
 endfor

 {using horizontal optical links, to get overall prefix sum}
 for all $P_{ij}$, where $2 \le i \le n$ and $1 \le j \le n/2$ do

      p-sum ⇐ [i, (j+n/2)]p_sum
      prefix ← prefix + p_sum
endfor
```

There are three phases in the algorithm. In the first phase, row-wise prefix sum is calculated. When the first phase is over, the last node in each row contains the sum of all the elements of that row. So, in the second phase we calculate prefix sum for these nodes. This sum is added to the row prefix sum in all the nodes of the respective rows to get the overall prefix sum (third phase). Each of the three phases require, n/2 moves. Therefore, the prefix sum of n^2 elements can be calculated in 1.5n time which is comparable to other traditional algorithms [11–13].

4 Conclusion

The proposed two-dimensional mesh network with optical interconnects reduces the diameter of the mesh topology and also improves its bisection width, while providing high-speed direct connections between nodes that are far apart. It has also been shown that the parallel algorithms are faster in comparison to the simple two-dimensional mesh topology.

References

1. Alverson, R., et al.: The Tera Computer System. In: International Conference on Supercomputing, Assoc. of Comput. Machinery (1990) 1–6
2. http://ed-thelen.org/comp-hist/intel-paragon.html
3. Scott, S., Thorson, G.: Optimized routing in the Cray T3D. In: First International workshop on Parallel Computer Routing and Communication, LNCS, Vol. 853 (1994) 281–294
4. Gara, N.A., et al.: Overview of the Blue Gene/L system architecture, IBM J. Res. & Dev., Vol. 49, No. 213 (2005) 195–212
5. Das, D., Sinha, B.P.: A new network topologies with multiple meshes. IEEE Transaction on Computers, Vol. 44, No. 5 (1999) 536–551
6. Chen, W.M., Chen, G.H., Hsu, D.F.: Combinatorial properties of mesh of trees,. In: International Symposium on Parallel, Architectures, Algorithms and Networks (2000) 134–139
7. Jana, P.K.: Multi-mesh of trees with its parallel algorithms. Journal of System Architecture, Vol. 50 (2004) 193–206
8. Marsden, G.C., Marchand, P.J., Harvey, P., Esener, S.C.: Optical Transpose Interconnection System Architecture, Optical Letters, Vol. 18, No. 3 (1993) 1083–1085
9. Wang, C.F., Sahani, S.: Basic Operation on OTIS-Mesh Optoelectronics Computer. IEEE Transaction on Parallel and Distributed Systems, Vol. 19, No. 12 (1998) 1226–1233
10. Sinha, B.P., Banyopadhyay, S.: OMULT: An Optical Interconnection System for Parallel Computing. LNCS, Vol. 3149 (2004) 302–312
11. Ladner, R.E., Fischer, M.J.: Parallel Prefix Computation, Journal of the Association of Computing Machinery, Vol. 27, No. 4 (1980) 831–838
12. Egecioglu, O., Srinivasan, A.: Optimal Parallel Prefix on Mesh Architecture, Parallel Algorithms Appl., Vol. 1 (1993) 191–209
13. Jha, S.K.: An Improved parallel Prefix Computation on 2-D Mesh Network, Procedia Technology, Vol. 10 (2013) 919–926

Enhanced TCP NCE: A Modified Non-Congestion Events Detection, Differentiation and Reaction to Improve the End-to-End Performance Over MANET

J. Govindarajan, N. Vibhurani and G. Kousalya

Abstract The characteristics of Mobile Ad hoc Network (MANET) like error and reordering degrades the performance of TCP-based applications. Among the many proposals to reduce the impact of non-congestion events, TCP-NCE has been designed as the unified solution to discriminate between non-congestion and congestion events, and to respond to the events. Our initial analysis on TCP-NCE and other schemes (TCP-DCR and SACK-TCP) showed that the existing schemes including TCP-NCE fail to improve end-to-end performance in the presence of congestion, error, and reordering due to mobility and multipath routing. To overcome this problem, we designed "Enhanced TCP NCE" protocol to reduce the false differentiation on non-congestion events and to optimize the response procedure to those events. Our simulation results showed that the enhancement increased the performance by 15–20% over TCP-NCE. In addition, the consistency in yielding the higher performance throughout the simulation is observed for our protocol.

Keywords MANET · TCP · Non-Congestion · Reordering · Error

J. Govindarajan (✉)
Department of Computer Science and Engineering, Amrita School of Engineering,
Amrita Vishwa Vidyapeetham, Amrita University, Coimbatore, India
e-mail: j_govindarajan@cb.amrita.edu

N. Vibhurani
Amrita Vishwa Vidyapeetham, Amrita University, Coimbatore, India

G. Kousalya
Coimbatore Institute of Technology, Coimbatore, India

P.K. Sa et al. (eds.), *Progress in Intelligent Computing Techniques: Theory, Practice, and Applications*, Advances in Intelligent Systems and Computing 719, DOI 10.1007/978-981-10-3376-6_48

1 Introduction

Transmission Control Protocol (TCP) is a flow controlled and congestion controlled end-to-end transport protocol to achieve the reliable transmission over unreliable network. The sending rate is self-controlled by acknowledgement (ACK) receiving rate and it is specified using the variable "Congestion Window" (CWND). Sender opens its transmission with a minimum sending rate of 1 MSS per RTT (i.e., cwnd = 1MSS), to avoid the early congestion and enters into slow-start phase. In this phase, for each new ACK, sender increases its CWND by one until CWND reaches the slow-start threshold (ssthresh) and hence sending rate is increased exponentially. After reaching ssthresh, it enters into congestion avoidance phase to avoid late congestion. In this phase, sending rate is increased linearly. The receiver sends new ACK if it receives the packet with expected sequence number. Otherwise, it sends duplicate ACK (dupack) to indicate the packet loss. Sender learns the congestion losses either from three dupacks (3dupacks) or timeout. Arrival of third dupacks at sender is the indication of weak congestion inside the network and hence sender reduces CWND by half. Then, retransmits the packets and enters into fast retransmit and fast recovery phase to maintain the sending rate during recovery. In high congestion case, the sender may not receive any ACK and hence timeout event will be invoked. It responds to high congestion by reentering into the slow-start phase.

In MANET, the sender receives 3dupacks in the case of congestion events or non-congestion events such as bit corruption, environment interferences, and packet reordering inside the network. Packet reordering happens due to multipath routing, disconnection of links due to node's mobility and packet retransmissions. TCP Sender which is running at endpoints considers the packet reordering as congestion loss and hence it retransmits the packets and reduces CWND unnecessarily. This packet reordering problem primarily impacts the congestion window growth and end-to-end performance degrades. Many TCP variants were proposed by the researchers which either delay or revoke the congestion response without differentiating between the packet loss and packet reordering. TCP-NCE is latest version of basic TCP which helps the sender to discriminate the losses. Our initial analysis showed that the inaccuracy of protocol in differentiating the reordering from error. In this paper, we recommend a method which helps the sender to respond to losses precisely and to take earlier action. To validate our protocol, we evaluated the performance of our proposal and the existing variants (TCP-DCR, SACK-TCP, and TCP-NCE) which address the reordering.

The rest part of the paper is organized as follows. Section 2 summarizes the related works of reordering problem. In Sect. 3, we presented our simulation and theoretical analysis of TCP-NCE. Section 4 describes our proposed "Enhanced TCP-NCE" scheme. Section 5 details the evaluated performance comparison of our protocol with others variants. Finally, in Sect. 6, we conclude this work with future direction to improve TCP performance.

2 Related Work

The existing solutions which address the reordering problem at transport level can be classified based on the nature of the algorithm. Initially, the variants were designed with the aim of avoiding spurious transmission like additional delay after receiving third dupack of TCP-DCR [1], dynamic duplicate acknowledgment threshold (dupthresh) of RR-TCP [2] and disabling congestion action of TCP-DOOR [3]. The variants which were proposed to address the other problems of TCP like TCP-Westhood [4], SACK-TCP [5] had been considered as solution to the reordering. Later, the researchers designed the variants with loss discrimination procedures to differentiate the non-congestion events like bit errors from congestion events like buffer overflow. Estimation of bottleneck link bandwidth, using calculated duplicate threshold for number of duplicate ACKs received, measuring the variation of inter-arrival time of packets are the few ideas which are followed by these procedures. TCP-NCE [6] is a recent version of TCP with Loss Differentiation Algorithm (LDA).

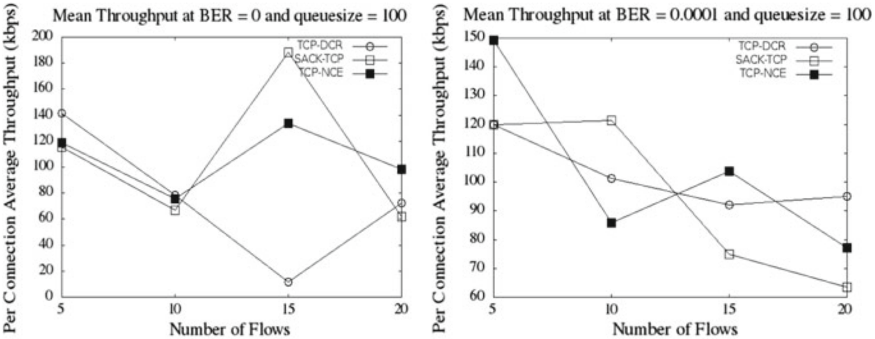
The authors of TCP-Delayed Congestion Response (TCP-DCR) addressed the non-congestion events problem by adding delay procedure with congestion response. In this procedure, sender delays the congestion action after receiving three dupacks. During this additional delay period the receiver may send the cumulative ACK after receiving the reordered packets. To maintain the same sending rate during the delay period, sender sends one new segment for every duplicate ACK. This solution delays congestion action without any loss classification.

TCP-NCE is designed with a loss discriminate algorithm to avoid unnecessary retransmissions by differentiating non-congestion losses from congestion losses, i.e., error and reordering from congestion events. To detect and differentiate the losses it uses queue length and flight size information (i.e., number of outstanding packets which are in transmission), respectively. It uses timestamp of TCP header to calculate the queue length. The flight size is considered as delay threshold. A TCP-NCE sender considers the packet losses as congestion when the current queue length is greater than the threshold value (Th-Val). Otherwise, it will be considered as non-congestion event. In the case of non-congestion event, it sends one new packet without any reduction in CWND and ssthresh, and calculates delay threshold (delay-thresh) which will be used to schedule the retransmission. The packet will be retransmitted after the expiration of delay threshold. When number of additional dupacks is greater than or equal to delay-thresh, the sender concludes the packet loss as error and hence retransmits the packet immediately. Otherwise, non-congestion event is concluded as reordering and hence the sender will resume its transmission of new data packets.

In [6] the authors of TCP-NCE proved that the protocol outperforms other variants (RR-TCP, TCP-PR, TCP-DOOR, TCP-CERL, and TCP-VENO). In our previous analysis [7], we have done detailed study on TCP-DCR, SACK-TCP, RR-TCP, and TCP-Westwood in different scenarios (Reordering with low/high congestion and no/low/high error). Two main causes of reordering (i.e., Multipath routing and Mobility) were considered in the analysis. Finally, we concluded that TCP-DCR and SACK-TCP outperform in low congestion with no or low error and fails in high congestion with high error. Since our previous analysis was among the variants without LDA, we believed that variant with LDA like TCP-NCE improves the performance in worst case, i.e., reordering with high congestion and high error. Hence, we have considered TCP-NCE for analysis and the enhanced version of TCP-NCE has been proposed in this work.

3 Analysis of TCP-NCE Over Mobile Ad Hoc Network (MANET)

We described the simulation setup which is considered to analyze the performance of TCP variants in Sect. 5. In simulation, mobile nodes were configured with MP-OLSR (Multipath-Optimized Link State routing) [8] with queue size of 100 packets to simulate the multipath forwarding behavior and low congestion in ad hoc network, respectively. Multipath forwarding is one of the main causes for packet reordering. To test the efficiency of TCP-NCE, experiments without error and with error (BER of 0.0001) were carried out. From the results, we observed that TCP-NCE behaves similar to other variants TCP-DCR and SACK-TCP in both cases as shown in Fig. 1. Except the result for the experiment with the numbers of flows equals to 15 and zero BER, a small variation (maximum of 20kbps) between TCP-NCE and other protocols can be observed in all other experiments. Also, we can observe that TCP-NCE does not yield consistent performance. For example, TCP-NCE shows higher performance in the experiment with "Number of flows" equals to 5, BER = 0.0001 and Queue Size = 100, and shows lower performance when the number of flows is increased to 10.



(a) Low congestion and no error case (b) Low congestion and error case

Fig. 1 Performance of TCP variants in presence of 2–4% reordering rate

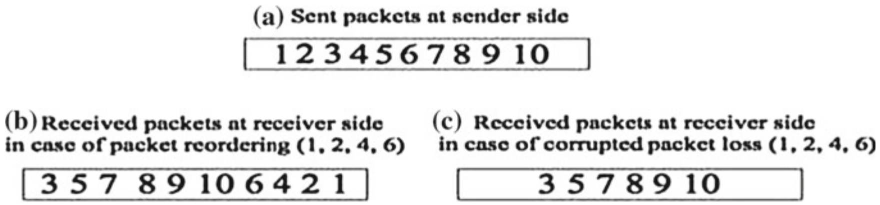

(a) Sent packets at sender side
1 2 3 4 5 6 7 8 9 10
(b) Received packets at receiver side in case of packet reordering (1, 2, 4, 6)
3 5 7 8 9 10 6 4 2 1
(c) Received packets at receiver side in case of corrupted packet loss (1, 2, 4, 6)
3 5 7 8 9 10

Fig. 2 Packet reordering and packet corruption scenarios

When multiple packets of same window corrupted and reordered, TCP-NCE tries to retransmit the unacknowledged packets upon threshold expire and fails to take immediate response upon packet corruption. Even though, TCP-NCE differentiates the non-congestion loss from reordering, the delayed response taken upon high number of reordered packets leads to unnecessarily retransmission.

Theoritical analysis on TCP-NCE protocol
Consider a scenario where sender sends 10 packets back-to-back (Fig. 2a). Suppose few packets of this window are reordered and received in the order as mentioned in Fig. 2b. Suppose few packets of this window are corrupted in transmission and the rest of the packets will reach the receiver as mentioned in Fig. 2c. For reordering case in Fig. 2b and packet error case in Fig. 2c, TCP-NCE concludes the losses as non-congestion loss while receiving the third dupack and computes the delay threshold as 9. Initially, non-congestion loss is considered due to packet reordering. Then, for each additionally received dupack, sender sends a new data packet to the receiver until it receives the nineth dupack. For reordering case in Fig. 2b, since the packets are only reordered without loss, sender will receive nine dupACKs from the receiver. After receiving the nineth dupack from receiver, the sender will designate the reordering as error and it retransmits the packet 1. Here, discrimination procedure of TCP-NCE gives false conclusion on reordering, i.e., reordering is concluded as error and false action will be performed. For packet error case in Fig. 2c, since the packets are dropped due to error, the sender will receive dupacks less than 9. Hence, TCP-NCE designates losses as reordering and increment the CWND without retransmitting the lost packets. In summary, TCP-NCE invokes the retransmission unnecessarily in reordering and causes for multiple cycles of retransmissions when multiple packets are dropped in same window.

4 The Proposed Solution: Enhanced TCP-NCE

Our enhanced version of TCP-NCE which addresses the reordering problem is discussed in this section. As mentioned in the previous section, to avoid false detection and action at the sender side a modification has been incorporated in the

TCP-NCE. As per this modification, initially (i.e., after receiving three dupacks) the sender considers the loss as error. For each duplicate ACK, CWND will be incremented to compensate the loss due to error. For each out-of-sequence packets the receiver will attach the SACK option header along with duplicate ACK and it will be used by the sender to update the “Delivered Packet List”. “Delivered Packet List” is a list of sequence numbers which are delivered to receiver and acknowledged by receiver either through ACK field or SACK option header. If the sequence number of SACK option of the received packet is less than the sequence number of last packet in the delivered list, then the sender prepares the “Reordering Window” and from that it will prepare the “Corrupted Packet List” for retransmission. Reordering Window” and “Corrupted Packet List” represent the list of sequence numbers which are reordered inside the network and the list packets are corrupted during the transmission inside the network respectively. Since the sender only transmits corrupted packet, we refer “Corrupted Packet List” also as “Missing Packet List”. When the number of dupacks received at sender exceeds the threshold (80% of CWND), the losses will be classified as reordering. When the timeout procedure is invoked the sender will prepare the “Corrupted Packet List” and retransmit those packets. To avoid the duplicate retransmissions sender maintains the retransmitted packets in “Retransmit Packet List”. Hence, the prepared “Missing Packet List” will cross verify with the past “Retransmit Packet List” and drops the packets which are already retransmitted This mechanism of responding to non-congestion events is shown in Algorithm 1.

TCP-NCE may fail to differentiate between congestion and non-congestion events when bottleneck link bandwidth information is not available at end nodes. As an enhancement we replace the differentiation procedure of TCP-NCE with loss discrimination procedure of Modified XCP [9] where the losses are differentiated by computing the queuing delay. In Modified XCP, the sender maintains the current estimated RTT as $\mathrm{RTT}_{\mathrm{cur}}$, the minimum of RTT as $\mathrm{RTT}_{\mathrm{min}}$ and the maximum RTT as $\mathrm{RTT}_{\mathrm{max}}$ for each packet transmission between the sender and the receiver. The queuing delay can be estimated using Eqs. (1) and (2). The ratio of queuing delay to $\mathrm{queuingdelay}_{\mathrm{max}}$ is compared with a threshold (λ). If queuing delay to $\mathrm{queuingdelay}_{\mathrm{max}}$ is less than the threshold (λ) (default value is 0.5), the packet losses are considered as congestion losses. Otherwise, the packet losses are considered as bit error losses.

$$\mathrm{queuingdelay} = \mathrm{RTT}_{\mathrm{cur}} - \mathrm{RTT}_{\mathrm{min}} \tag{1}$$

$$\mathrm{queuingdelay}_{\mathrm{max}} = \mathrm{RTT}_{\mathrm{max}} - \mathrm{RTT}_{\mathrm{min}} \tag{2}$$

Algorithm 1: Loss Discrimination Algorithm of Enhanced TCP-NCE

```
Initialization:
 loss ← false, missing_packetlist ← null, $\lambda \leftarrow 0.5$
On receiving each DupACK:
begin
 dupack ← dupack +1
 if dupack == 3 then
   if queuingdelay <= $\lambda$ x $queuingdelay_{max}$ then
        loss ← true, noncong_loss ← true, cong_thresh ← 0.80 * CWND
   else
       loss ← true,   cong_loss ← true,   Congestion_action()
  end
end
if dupack > 3 and noncong_loss == true then
     Add seqnum(SACK header) to delivered packet list
     CWND ← CWND + 1 MSS
      if dupack == noncong_thresh then
        missing_packetlist ← Update_missing _pktlist(()
     end
    else if dupack < noncong_thresh then
        if seqnum(SACK header) < last_seqnum_deliveredpktlist() then
        reodering_event ← true
        missing_packetlist ← Update_missing _pktlist()
         retransmit pkts(missing_packetlist)
       else
            send_newpkt()
        end
    end
end
On Timeout:
 begin
   missing_packetlist ← Prepare_missing _pktlist(()
   retransmit pkts(missing_packetlist)
end
```

5 Performance Analysis of Enhanced TCP-NCE

To study performance of proposed "Enhanced TCP-NCE", we considered existing TCP-NCE, TCP-DCR, and SACK-TCP, and the results are summarized in this section. The experimental setup which was used in our previous work [7] replicated here. Figure 3 shows the topology which is considered for our analysis. We considered two different scenarios to represent the different levels of reordering. The first scenario was designed with low congestion by setting the queue size as 100 and BER as 0.0001. In second scenario, the high reordering is simulated by setting

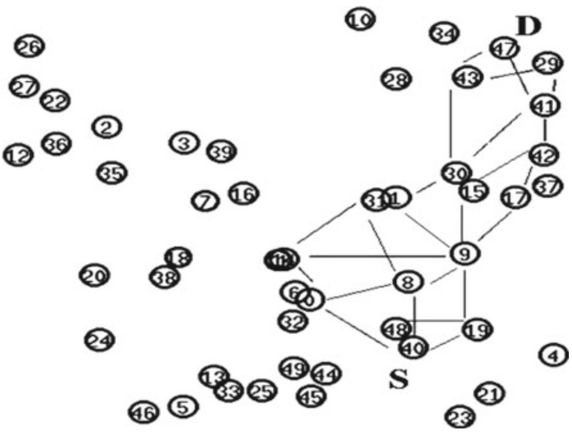


Fig. 3 A Topology with 50 nodes (The paths between one pair of nodes are drawn)

Table 1 Simulation parameters

Parameter	Value
Simulation duration	1000 s
Simulation area	800 m × 800 m
Number of connections	5,10,15,20
Queue size (Packets)	20,100
Error rate	0.0001
Mobility model	Gauss-Markov mobility model
Transport level protocols	TCP-DCR/SACK-TCP/TCP-NCE/EnhancedTCP-NCE
Routing protocol	MP-OLSR (Multipath-Optimized Routing Protocol)

queue size as 20 and BER as 0.0001. Table 1 summarizes our simulation parameters.

Scenario 1*: *Low Congestion with High Error (queue size = 100 and BER = 0.0001)

The simulation results of low congestion and high error are presented in Fig. 4a, b. In this scenario, “Enhanced TCP-NCE” achieves 10–20% higher throughput and goodput than TCP-NCE and TCP-DCR. SACK-TCP fails to sustain its performance while increasing the number of connections over the network. Due to the absence of reordering detection procedure, TCP-DCR and SACK-TCP yield low performance (max. of 50 kbps) compared to “Enhanced TCP-NCE” (120 kbps). In contrast to TCP-NCE, our Enhanced TCP-NCE reduces the number of false loss discriminations and invokes earlier retransmission of corrupted packets. Hence, the proposed scheme achieves higher performance. Designing the reliable protocol (i.e., optimized higher throughput in all cases) is the main challenge of a protocol designer. From Fig. 4, we can see that the “Enhanced TCP-NCE” yields throughput of 120 kbps and maintains its performance for different number of connections while the existing protocols fail to provide reliability. Other than the packet

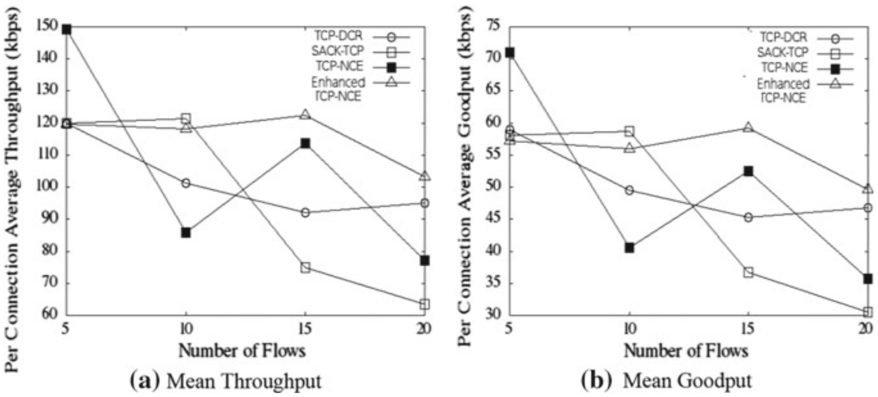


Fig. 4 Performance of the connections with medium reordering at BER = 0.0001 and Queue size = 100 packets

reordering due to the path change, the frequent unnecessary retransmissions also causes the packet reordering. Our measurements have shown that most of the connections (8 out of 10, 16 out of 20) falls in the medium reordering category when the "Enhanced TCP-NCE" is used, i.e., it is avoided the chain of reordering. When TCP-NCE is used at transport level, most of connections fall under higher reordering. This shows that our modification avoids unnecessary retransmissions and hence frequent reordering is avoided.

Scenario 2*: *High Congestion and High error (queue size = 20, BER = 0.0001) The high congestion and error causes the frequent retransmissions and reordering. Figure 5 shows that "Enhanced TCP-NCE" achieved 10–15% performance gain in terms of both throughput and goodput compared to the existing TCP variants. In case of timeout, Enhanced TCP-NCE retransmits the unacknowledged packets

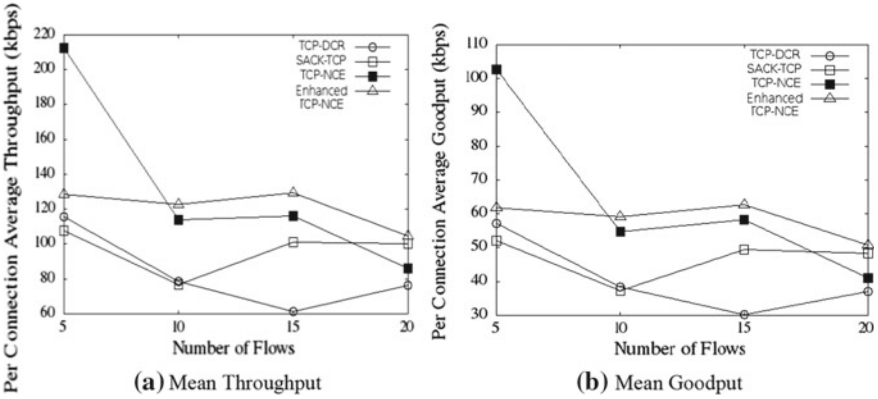


Fig. 5 Performance of the connections with medium reordering at BER = 0.0001 and Queue size = 20 packets

without reducing the CWND, i.e., the sender maintains the sending rate of 120 kbps and hence increases the utilization when number of connections is increased. The discrimination of non-congestion from congestion loss using queue delay calculation and discrimination of reordering from error of our proposed scheme helps the sender to avoid false actions like unnecessary retransmission, unnecessary reduction in sending rate, and optimizes the time to recover the multiple packet losses in same window.

Comparing efficiency of TCP variants

The comparison of efficiency (%) (Refer Eq. 3) of TCP protocols is shown in Fig. 6. In Eq. (3), "Transmitted bytes" refers the total number of TCP payload bytes transmitted which includes both original bytes and retransmitted bytes. If a protocol optimizes the number of retransmissions and avoided the spurious retransmissions, then the efficiency of the protocol will increase. The protocol with higher efficiency sometimes yields low throughput. Hence, in our analysis, we studied the efficiency of protocols along with their achieved throughput. To increase the accuracy of our study, we calculated the efficiency in best case, average case and worst case i.e. max-efficiency, average efficiency and minimum-efficiency.

$$\text{Efficiency} = \frac{\text{Transmitted bytes} - \text{Retransmitted bytes}}{\text{Transmitted bytes}} \times 100 \qquad (3)$$

The result of our study on efficiency and throughput is shown in Fig. 6 for the scenario 2 (high congestion and high error). Figure 6 shows that "Enhanced TCP-NCE" achieves 5% higher efficiency than TCP-NCE and TCP-DCR in average efficiency. Since the existing SACK-TCP and our "Enhanced TCP-NCE" are designed based on SACK, both protocols avoids spurious retransmission and hence efficiency of the protocols is same. Even while comparing the Minimum and

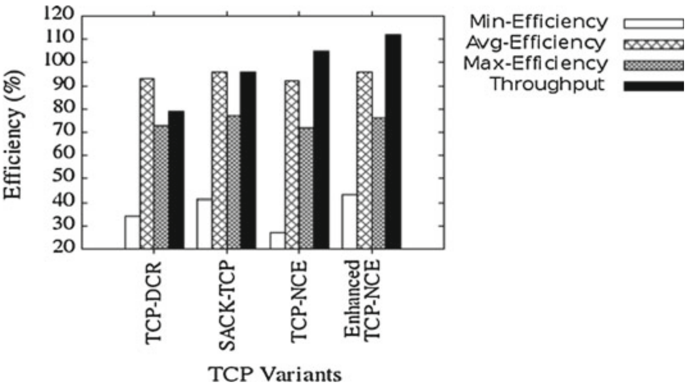


Fig. 6 Comparison of TCP variants in terms of efficiency in case of high reordering due to both congestion and non-congestion loss (queue size = 20, BER = 0.0001 and no. of connections = 20)

Maximum efficiencies of variants, Enhanced TCP-NCE and SACK-TCP achieves higher efficiency (40% in case of Min. efficiency and 75% in case of Max. efficiency) than others. This shows that both variants avoid the unnecessary retransmission in case of high reordering. However, SACK-TCP fails to achieve higher throughput due to the absence of loss discrimination mechanism and compensation procedures for error. Due to space constraint, we presented efficiency plot when number of connections is 5. For other cases, the same observation is made with a significant difference between SACK-TCP and "Enhanced TCP-NCE" in throughput.

6 Conclusion

The occurrence of congestion and non-congestion events (error and packet reordering) in MANET causes the TCP sender to take false decision which leads to the poor performance. TCP-NCE is an end-to-end solution designed with loss discrimination procedure for congestion, error, and reordering. In this work, we have modified the TCP-NCE to avoid false classification on non-congestion events (error and reordering). Our simulation results have proved that the enhanced TCP-NCE optimizes the end-to-end performance, i.e., increases the throughput by 15% and consistency in yielding higher performance. Since in this work we concentrated on redesign of non-congestion event classification to its increase accuracy, as a minor change we replaced congestion detection mechanism of TCP-NCE with Loss Discrimination Algorithm (LDA) algorithm of MXCP. Further investigations are required on LDA algorithms for the detection of congestion events.

References

1. Bhandarkar S, Reddy ALN (2004) TCP-DCR: Making TCP robust to non-congestion events. In: Mitrou, N., Kontovasilis, K.P., Rouskas, G.N., Iliadis, I., Merakos, L.F. (eds.) NETWORKING 2004. LNCS, vol 3042. Springer, Heidelberg, pp. 712–724
2. Zhang M, Karp B, Floyd S, Peterson L (2003) RR-TCP: A Reordering-Robust TCP with DSACK. In: Proc. The Eleventh IEEE International Conference on Networking Protocols (ICNP 2003), Atlanta, GA, November 2003, pp. 95–106
3. Wang F, Zhang Y (2002) Improving TCP performance over mobile Ad-Hoc networks with out-of-order detection and response. In: ACM MOBIHOC,Vol 9–11,Lausanne, Swizerland, pp. 217-225
4. Casetti C, Gerla M, Mascolo S, Sanadidi M, Wang R (2002) TCP Westwood: End-to-end congestion control for wired/wireless networks. ACM/Kluwer Wireless Networks (WINET) Journal 8(5): 467–479
5. Fall K, Floyd S (1996) Simulation-based comparisons of Tahoe, Reno and SACK TCP. ACM SIGCOMM Computer Communication Review 26:5–21. doi:10.1145/235160.235162

6. Sreekumari P, Chung S (2011) TCP NCE: A unified solution for non-congestion events to improve the performance of TCP over wireless networks. EURASIP J Wirel Commun Netw 2011:23. doi:10.1186/1687-1499-2011-23
7. Govindarajan J, Vibhurani N, Kousalya G (2015) An Analysis on TCP Packet Reordering Problem in Mobile Ad-Hoc Network. Indian Journal of Science and Technology 8(16)
8. Yi J, Adnane A, David S, Parrein B (2011) Multipath optimized link state routing for mobile ad hoc networks. Ad Hoc Networks 9(1):28–47. doi:10.1016/j.adhoc.2010.04.007
9. Sun Y, Ji Z, Wang H (2012) A modified variant of eXplicit Control Protocol in satellite networks. Journal of Computational Information Systems 8(10):4355–4362

Intelligent Building Control Solution Using Wireless Sensor—Actuator Networking Framework

Anindita Mondal, Sagar Bose and Iti Saha Misra

Abstract Intelligent building refers to a residence that is automated through a network of electronic devices which cooperate transparently to provide protection and comfort to the residents and minimize the energy consumption drastically. In this paper, a novel approach is made to design an intelligent building control solution using wireless sensor actuator networks (WSAN) in hardware domain. The intelligent building control solution provides automation to most household operations such as intrusion alarm, environment monitoring and controlling, asset tracking, etc. Real-time sensing is established with the help of sensor modules whereas real-time tracking is achieved with the deployment of active radio frequency identification (RFID) module. The WSAN greatly consists of gateway, routers, and end devices. The main controlling unit of WSAN is AVR microcontroller which manages the transceiver section and processes the received data accordingly. This paper focuses on remote monitoring and management system for a partly automated building equipped with sensors and actuators. Here, we also explore the prospects of wireless mesh networks and design a framework for developing an intelligent and smart environment.

Keywords Wireless sensor actuator networks (WSAN) · Wireless personal area network (WPAN) · IEEE 802.15.4 · Zigbee · Active RFID · Wireless mesh network (WMN)

A. Mondal (✉) · S. Bose · I.S. Misra
Department of Electronics and Telecommunication, Jadavpur University, Kolkata, India
e-mail: anindita.mishti@gmail.com

S. Bose
e-mail: sgrbose@gmail.com

I.S. Misra
e-mail: itisahamisra@yahoo.co.in

P.K. Sa et al. (eds.), *Progress in Intelligent Computing Techniques: Theory, Practice, and Applications*, Advances in Intelligent Systems and Computing 719, DOI 10.1007/978-981-10-3376-6_49

1 Introduction

Recently, the pledge of many countries to cut the annual consumption of primary energy has increased the demand of products that reduce the energy consumption [1]. This acted as a catalyst for designing and developing an intelligent building control solution which provides energy efficient strategies with microcontroller-driven transceiver to minimize power consumption.

The primary objective of this paper is to design the hardware platform for the intelligent building control solution using microcontroller and ZigBee-based RFID module. Visualization software has been developed which offers real-time network monitoring, alert notification, and reporting functionalities to manage mobile objects and environment from a single, scalable, unified platform.

The AVR microcontroller is the main component of the hardware circuitry which controls the transceiver module, i.e., ZigBee and is interfaced with various sensors to obtain different measuring parameters such as light, temperature, humidity, etc. A closed-loop control system has been developed for the automation of different electric appliances inside the building with the deployment of actuators. The sensor data are collected and updated in the server periodically to obtain environmental statistics. Simultaneously active RFID router locates the position of a particular object strategically.

1.1 Active RFID-Based Integrated Tracking and Sensing System

Radio frequency identification (RFID) is a device (typically referred to as an RFID tag) that can be affixed/mounted on a product, animal, or person for the purpose of unique identification and tracking using radio waves. There are generally two types of RFID tags, i.e., active RFID and passive RFID. Active RFID uses an internal power source (battery) within the tag to continuously power the tag and its RF communication circuitry, whereas passive RFID relies on RF energy transferred from the reader to the tag to power the tag [2, 3]. Passive RFID requires stronger signals from the reader, and the signal strength returned from the tag is constrained to very low levels. Active RFID allows very low-level signals to be received by the tag (because the reader does not need to power the tag), and the tag can generate high-level signals back to the reader. In our system, we make use of active RFID compliant with IEEE 802.15.4 modeled wireless networking.

In this scenario, each RFID tag has been attached with a particular asset to locate its position and if the asset is displaced from its position an alarm will be generated and the location of the asset will be displayed in the GUI.

2 Architecture of Real-Time Tracking and Sensing System Using IEEE 802.15.4 Based Wireless Mesh Network

In this scenario, an infrastructure is designed where routers and gateways are backbone of the infrastructure. RFID tag sends sensor data to the gateway via router. While sending the data, the destination address is kept fixed whereas intermediate hopping address varies accordingly. Sensor data received at the gateway are analyzed, processed, and actuation based on the processed data is initiated. The data for actuation is sent back to respective node via routers again to achieve the desired environmental parameters. In this way architecture for real-time sensing system has been established.

In real-time tracking system, the RFID module sends beacon packet to the nearest router then to the gateway via other routers. Inside the beacon packet the source address and destination address are incorporated. The source address gets extracted from the received packet at the gateway. Source address provides the information about the nearest router the tags are attached to and maps the position of the tag in the GUI. The location of the tag is updated after certain interval of time. If there is any undesired movement, alarm with actuation (e.g., automatically door lock) will be generated (Fig. 1).

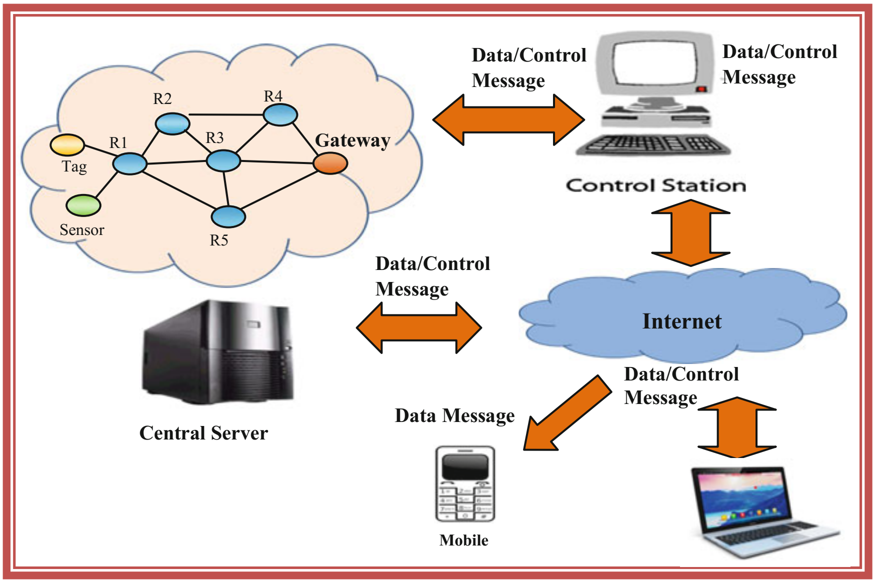


Fig. 1 Architecture of real-time tracking and sensing system

3 Hardware Description of Intelligent Building Control Solution

A single hardware module, i.e., wireless sensor actuator mote (WSAM) has been designed in such a way that it can be configured as coordinator, router, or end device. The firmware programming embedded inside the microcontroller is different for coordinator, router, and end device (Figs. 2 and 3).

3.1 Block Diagram of WSAM

WSAM consist of AVR Microcontroller, FT232RL, voltage controller, solid state relay, diode, resistor, potentiometer, capacitor, power supply socket, etc. The power supply provided is either by 9–12 V DC adapter or with rechargeable battery. The

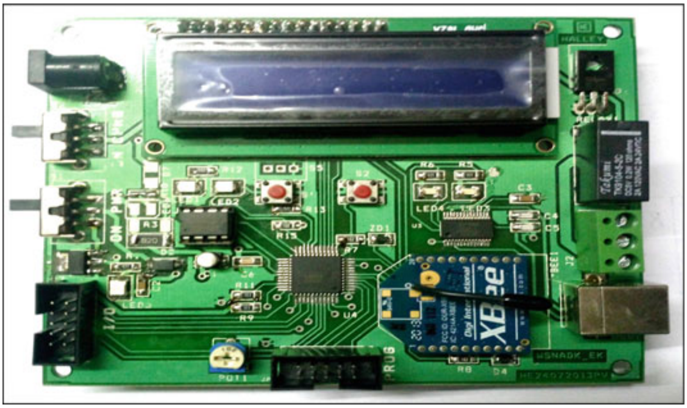


Fig. 2 Outlook of wireless sensor actuator mote

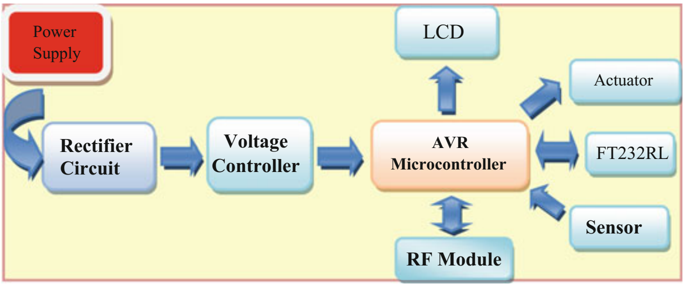


Fig. 3 Block diagram of WSAM

voltage controller 7805 IC is used for 5 V supply to the microcontroller then the rectifier circuit is connected to bypass AC (alternating current) followed by a diode to restrict the flow of current in opposite direction. The FT232RL is a USB to TTL serial converter IC used in applications where USART devices need to communicate to external devices through USB. Sensors are interfaced with the ADC (Analog to Digital Converter) pin of the microcontroller where the 10-bit digital value of sensor data is calibrated and sent to the gateway. Actuators such as buzzer, led, and relay are connected to the output pin of microcontroller. The gateway module receives the data from router, updates it in the server, and sends back the actuation to the required node.

The platform used to compile embedded C code is mikroC PRO for AVR. After successful compilation of the C code corresponding HEX file is generated to burn the microcontroller [4]. Extreme Burner software [5] transfers the HEX file from personal computer to the microcontroller with the help of AVR programmer.

The visualization software is created by JAVA programming and updating of data in the server is done through MySQL [6]. The integration of WSAN, visualization software, and database management is a rigorous process. Time synchronization is one of the key issues to be managed carefully.

3.2 Features of the Prototype and Description of Sensors Used

We have developed a prototype with the following features.

(a) Ambient environment monitoring (monitoring of temperature and humidity)
(b) Lighting control
(c) Liquefied petroleum gas (LPG) gas leakage detection
(d) Fire and smoke detection and alert generation
(e) Wireless intrusion detection system
(f) Data logging in the web server
(g) Generation of alarms

The following sensors are interfaced with the microcontroller.

3.1.1 Humidity Sensor: The HIH-5030 (from Honeywell) sensor is used as the sensor device for the detection of humidity. A humidity sensor is a device that measures the relative humidity of a given area. A humidity sensor can be used in both indoors and outdoors. The sensor gives analog voltage based on humidity presence. This is attached with an ADC channel of WSN board.

3.1.2 Light Sensor: The TLDR-7630 sensor is used as the sensor device for detection of light.

3.1.3 Passive Infrared Sensor: The 555-28027 (of Parallax) is used as passive infrared sensor to detect the presence of human.

3.1.4 Temperature Sensor: The MCP9700 uses as low-power voltage output temperature sensor from microchip. The range of the sensor is 0–70 °C which is sufficient for measuring ambient temperature [7].
3.1.5 Smoke Sensor: We use First Alert SA340 smoke sensor. This sensor uses ionization technology which is the best for detecting flames and smoke.
3.1.6 LPG Sensor: The sensor measures any leakage of LPG gas from cylinder. We use MQ-6. It has high sensitivity to Propane, Butane, and LPG, it also responds to natural gas.

4 Wireless Routing of the Information

We have chosen IEEE 802.15.4-based wireless mesh network [8] as the wireless backbone for communicating the information with a remote station. Tags, routers, and gateway are used to capture the identity and status of tagged objects with sensor value. Tags, routers, and gateways are all low-power IEEE 802.15.4 compliant active RF devices. The routers are arranged in a networking topology called "mesh". Mesh network is a type of network where each node can communicate with multiple other nodes thus enabling better overall connectivity [9].

5 Field Implementation

We have deployed this system inside a building for testing.

A brief description of the deployment is as follows:

- We have deployed seven routers identified by R1, R2, R3, R4, R5, R6, and R7 which will form a wireless mesh network to relay sensor data with location information from five different sensor boards (tags with externally attachable sensor). Seven sensors are used here: temperature sensor, humidity sensor, light sensor, passive infrared sensor, smoke sensor, liquid petroleum gas sensor, and active infrared detector. We have used one coordinator which will be attached with a computer through a USB cable (Control Station Server) where the visualization software will run.
- We have programmed WSAM board to make it function as router (seven WSAM boards have been used for making seven routers) and also tag (five WSAM boards have been used for this purpose). Each tag can be integrated with multiple sensors.
- Five tags with sensor devices have been installed in different rooms.
- After the installation of the prototype system as mentioned above, all the devices like sensors, routers, and the coordinator have been powered up and the software starts running on the laptop (control station).

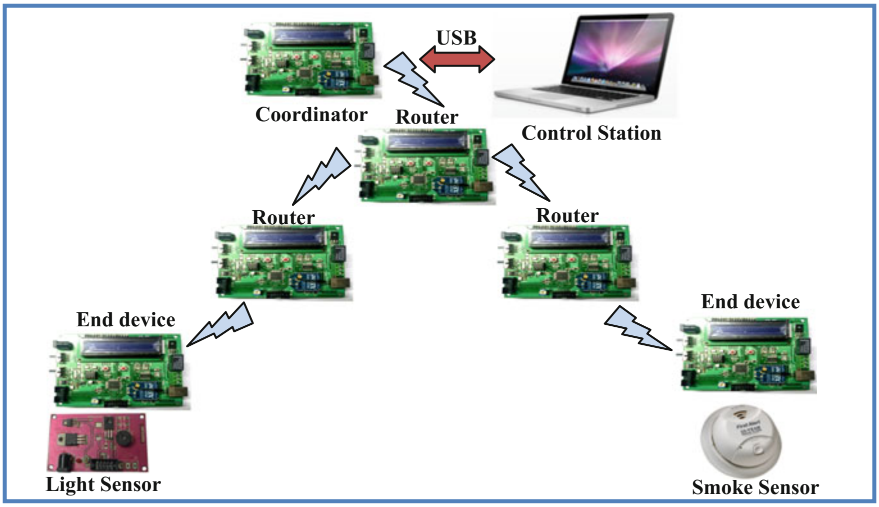


Fig. 4 Conceptual framework of deployment

- Sensor devices are powered up and start collecting the data from the sensors attached to them at specified intervals.
- The routers forward the transmitted data from the sensor devices to the coordinator at the monitoring station.
- The coordinator, as soon as it receives the sensor data from sensor device via the routers, populates the data into the data base. Before populating the data into the database software does calibration on received data.
- The web-based software keeps a record of all the data as and when received by the coordinator and also shows the live data on map-based visualization page.

The conceptual framework (Fig. 4) for the deployed system is given below. The framework contains one coordinator which is connected to the control station computer through USB cable, three routers, and two end devices attached with sensor module. Data packets from the sensor device are transmitted to the coordinator via multi-hop routing.

6 Real-Time Data Analysis

6.1 Calibrated Temperature and Humidity Report: We have done the field trial of our system continuously for 8 days. We obtained around 11,500 data from each sensor device during our survey. The day-wise report on temperature and humidity sensor is presented below.

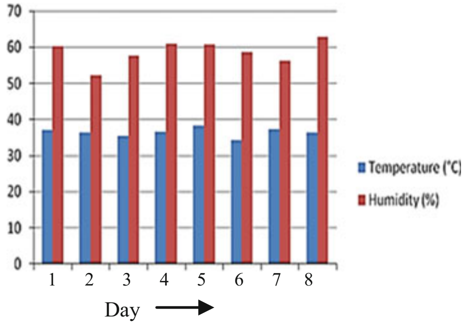


Fig. 5 Recorded temperature and humidity data

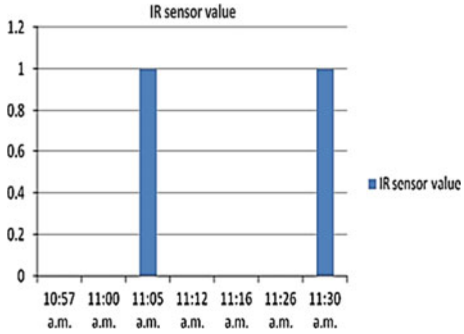
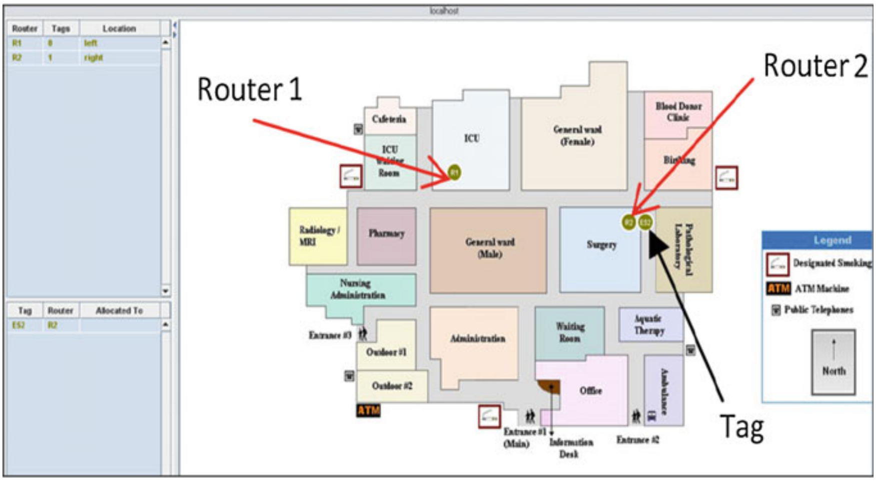


Fig. 6 IR sensor data

6.2 Intrusion Detection with the Help of Active Infrared Detector: We have deployed this intrusion detection system using IR sensor in a single day. Only two types of values are obtained here, either 0 or 1. If any intrusion is detected, the value obtained is 1, otherwise 0. In this paper, we have presented the report on intrusion detection using IR sensor in a single day (Figs. 5, 6 and 7).

7 Cost–Benefit

The emerging technology and cost of the product are pivotal issues to be considered to estimate the growth of the product in the market [10]. At present there are many building solutions available in the market, which are mainly concerned about intelligent controlling of electric appliances to decrease the power consumption. The price of these products is much higher than the WSAN designed in this paper. The integration of WSAN with periodical entry of measured data into the server and real-time tracking is the main significance of the prototype designed here. The cost of WSAN is about 60% less than the existing market product with more in-built features.



Router	Tags	Location
R1	0	left
R2	1	right

Tag	Router	Allocated To
ES2	R2	

Fig. 7 Screenshot of visualization software (showing position of tag and routers)

8 Conclusion

This paper presented the hardware design of WSAM and its field implementation for real-time tracking and sensing. The experimental work for the testing of the WSAM in real-time tracking and sensing was performed successfully. Results are obtained in the form of stored data in the server and actuation based on sensing and tracking. The prototype is checked for fault tolerance and necessary components are added to provide reliability and long life cycle. Limitation of the presented hardware design is constraint of I/O (Input/Output) pins available in the microcontroller and fully mesh network leads to flooding of data. Future scope of work is to minimize data congestion by firmware programming and optimization of the embedded software, and more efficient use of radio module by controlling the transmission power.

References

1. Hanne Grindvoll, Ovidiu Vermesan, Dr. Tracey Crosbie, Roy Bahr, Nashwan Dawood and Gian Marco Revel: *A Wireless Sensor Network For Intelligent Building Energy Management based On Multi Communication Standards–A Case Study*, Journal of Information Technology in Construction - ISSN 1874-4753, May 2012.
2. Xuhui Chen and Peiqiang Yu, *Research on Hierarchical Mobile Wireless Sensor Network Architecture with Mobile Sensor Nodes,* Proceedings of the IEEE 3rd International Conference on Biomedical Engineering and Informatics (BMEI 2010), pages 2863–2867.

3. Peter Corke, Tim Wark, Raja Jurdak, Wen Hu, Philip Valencia, and Darren Moore: *Environmental Wireless Sensor Networks,* Proceedings of the IEEE, Vol. 98, No. 11, November 2010.
4. [Online]. Available: http://www.mikroe.com/mikroc/avr/ide.
5. [Online]. Available: http://extremeelectronics.co.in/avr-tutorials/gui-software-for-usbasp-based-usb-avr-programmers.
6. [Online]. Available: https://www.mysql.com.
7. [Online]. Available: http://www.microchip.com/wwwproducts/MCP9700.
8. *Zigbee Specification,* Zigbee Alliance, June, 2005.
9. Wenqi (Wendy) Guo, Willam M. Healy and Mengchu Zhou, *Wireless Mesh Networks in Intelligent Building Automation Control: A Survey*, International Journal of Intelligent Control and Systems Vol. 16, No. 1, March 2011, 28–36.
10. Aamir Shaikh and Siraj Pathan, *Research on Wireless Sensor Network Technology,* International Journal of Information and Education Technology, Vol. 2, No. 5, October 2012.

Security Framework for Opportunistic Networks

Prashant Kumar, Naveen Chauhan and Narottam Chand

Abstract Opportunistic Networks have evolved as special class of mobile ad hoc and delay tolerant networks which have a vast range of applications. In opportunistic networks, the permanent links among the nodes are absent and delay is high. Due to self-organized nature of opportunistic networks, these networks have many security threats, e.g, how to protect the data confidentiality, integrity, privacy as well as the trust among the nodes. In this article, we present the specific security challenges to opportunistic networks and analyze related security requirement. Based on these discussions, we propose a general security framework for opportunistic networks and point out the future research direction in opportunistic networks.

Keywords Opportunistic networks · Security · Privacy · Security framework · Delay tolerant networks · Trust management

1 Introduction

Opportunistic Networks (OppNets) provide attractive solution in low connectivity regions. In OppNets, the information is exchanged between mobile devices when they encounter each other. There are a wide range of applications of OppNets—disaster and rescue networks [1], wildlife monitoring [2], social networking, e.g., DakNet [3], PodNet [4], to name a few. OppNets are the special class of ad hoc networks where end-to-end connectivity among the nodes is absent. Further routing mechanism in OppNets is receive–carry–forward instead of receive–forward [5].

P. Kumar (✉) · N. Chauhan · N. Chand
National Institute of Technology Hamirpur, Hamirpur, India
e-mail: prashantkumar32@gmail.com

N. Chauhan
e-mail: naveen@nith.ac.in

N. Chand
e-mail: nar@nith.ac.in

P.K. Sa et al. (eds.), *Progress in Intelligent Computing Techniques: Theory, Practice, and Applications*, Advances in Intelligent Systems and Computing 719, DOI 10.1007/978-981-10-3376-6_50

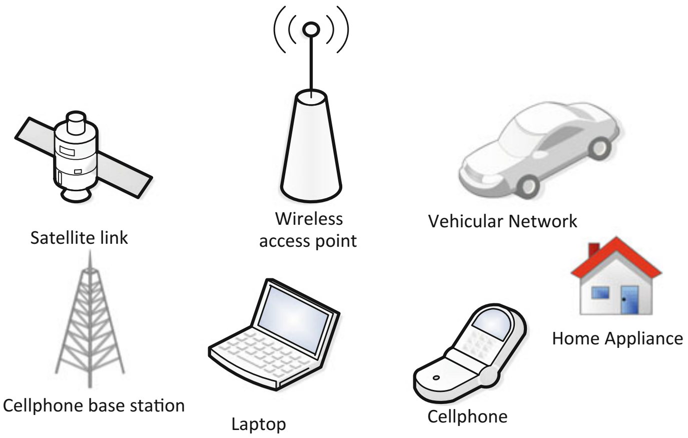


Fig. 1 Example of *devices* in OppNets

Due to intermittent connectivity, seed deployment, different routing mechanism, mobility, delay tolerance nature, OppNets have many security concerns. Security solutions designed for traditional wireless and ad hoc networks may not be directly applicable to OppNets, as OppNets own different and unique features.

A variety of devices with different communication technologies are used in OppNets as shown in Fig. 1. Thus devices used in this type of networks vary in communication range, computational resources, battery power, etc. In nutshell devices are heterogeneous in OppNets. Further mobility and frequent link disruption make OppNets topology highly dynamic and flexible. In the absence of direct link between the source and destination, the nodes have to hold the data with themselves till the next communication opportunity is found. Thus the data moves closer to destination hop by hop. If any node(s) is selfish then network performance may degraded by a huge factor. Selfish nodes are those nodes which forward their messages to relay nodes, but not wish to relay themselves. This fact implies the importance of trust and cooperation among the OppNets nodes. User privacy is another important issue in OppNets since context information is used in many of the OppNets routing protocols. This may be sensitive to some users.

2 Specific Security Challenges in OppNets

In this section, we describe specific and unique challenges concerned with designing security protocols and solution for OppNets.

1. **OppNets Seed Deployment**: OppNets grows from its *seed.* In the beginning set of nodes are employed together to build an OppNets. Then *seeds* invite other available devices to join the OppNets and network starts to grow [6]. Further

these nodes are able to self-localize as well as self-configurable according to network requirement. All these characteristics make OppNets so easy to get attacked.

2. **Mobility**: As nodes are highly mobile in OppNets, frequent disconnection occurs and a constant end-to-end connection never exists, so any security solution for OppNets must be adaptable to this extremely dynamic topology.
3. **Routing Mechanism**: OppNets uses store–carry–forward routing mechanism. This mechanism is helpful to extend the connectivity and to improve the data delivery. Due to this, it is very challenging to develop security solution for OppNets in the absence of end-to-end key management and a redesigning is needed for all established security algorithms.
4. **Heterogeneity**: In this era of ICT, there are plethora of devices, which are equipped with different radio interfaces that have variation in transmission and receiving capabilities and may operate over different frequency bands [7]. Each node may have a different hardware as well as software configuration, may cause to variation in processing capabilities. Further this might lead to naming and addressing problems [8].
5. **Decentralize Nature**: Heterogeneous networks are highly disconnected. They are decentralized in nature. Designing security protocols and solutions for this kind of network is complex.

3 Security Requirements

In this section, we will discuss the security requirements for OppNets. As in any network, the security requirements are *Privacy*, *Authentication*, *Integrity,* and *Non-repudiation*. In addition to these properties, *Trust* is a very important aspect and requirement in OppNets security. Collectively these properties can acronym as *PAINT* (***P**rivacy*, ***A**uthentication*, ***I**ntegrity*, ***N**on-repudiation,* and ***T**rust*). Thus we can say any security solution for OppNets must satisfy *PAINT* properties. Now we will discuss each attribute of *PAINT* properties.

1. **Privacy**: Privacy is twofold in OppNets. One is the privacy of data being exchanged, i.e, confidentiality of data should be maintained. Need to avoid the eavesdropping and tampering of data packets at any intermediate node in the network. To maintain privacy of data, the user needs to perform encryption of the data packed and key must be shared only with its destination. Second is the privacy of user. The privacy of user must be maintained as many of the routing algorithms use *context* information in order to find the suitable routes for routing.
2. **Availability**: In OppNets availability can be seen in two terms: availability of data and availability of node. Availability of data is existence of data to response some node, this can be good in OppNets because of distribution of data and can

be further enhanced by avoiding unnecessary dropping. Availability of nodes is forwarding data to the destination when the best possible relay node is available.
3. **Integrity**: Maintaining the consistency of data is termed as integrity of data in the network. In OppNets data is distributed and multiple copies exist, so integrity need exceeds and also becomes quite difficult. For keeping integrity one most prominent solution is to encrypt data and another may be the modification required by an authorized node and then to maintain a track of modification and advertise to neighbors.
4. **Non-repudiation**: Non-repudiation is that originator of data cannot deny it later. It can be done by providing authentication to the data packet with some digital signature of the source.
5. **Trust Management**: For secure communication, trust is the most vital parameter in the distributed scenarios. Trust is a factor which helps in validation of the legitimate user. Several attacks can be avoided by considering the thrust of a user. In OppNets trust can be build based on social interactions, frequent encounters, etc. Trust is a subjective property used for future interactions and based on past analysis.

4 Security Threats to OppNets

In this section, we will describe specific and unique challenges concerned with designing security protocols and solution for OppNets.

1. **Threats to Authenticity**: In OppNets the source and destination of nodes and other forwarding details are needed to be available to every relaying node for data forwarding to best possible node. This is an advantage for attacker to modify its source details. Known details of source can be used to inject false data to network with these details and misguide the network.
2. **Threats to Integrity**: As there are multiple number of copies that exist for data in the network and intruder changes the payload of packet at some places and the changed copy is forwarded to the destination, there is no way to find out that data is not original. This is due to the distribution of data in the network, so it can be reduced by limiting this distribution of data. One more solution to keep the integrity is the encryption of payload with only source and destination sharing the encryption/decryption mechanism.
3. **Denial of Service (DoS)**: For OppNets location of node plays a vital role and if it is misused then degradation of the network is enhanced. Response to legitimate nodes is denied due to certain reasons, i.e., fake location of any node leads to forwarding in wrong direction or no forwarding, Sybil nodes in the network, gray hole attack in which data is dropped and then retransmission demand.
4. **Confidentiality Risk**: Wireless signals are more prone to eavesdropping and in OppNets data is distributed also which increases its probability. Mobile nodes may be intercepted and tampering may be done by any malicious node. Also

data can be modified by the intermediate nodes. To avoid confidentiality risk, ensure data is not modified in the network and eavesdropping can be avoided by using the encryption schemes for encrypting the payload.

5. **Privacy Attack**: Privacy attacks can be classified in two categories: identity privacy and location privacy. As data is kept by intermediate nodes in OppNets which is easy to access for malicious nodes and privacy can be breached. In OppNets data forwarding is based on the destination node specified in data packet. So data packet is needed to be read by the relay nodes. The location of node should be kept private as any attacker may use the location of the node and inject a fake location of node by hiding its actual one, which leads to degrading the quality of forwarded data and data loss.

5 Proposed Security Framework

OppNets are based on the spontaneous response of the devices. Most of the time owners of the devices are unknown to each other, which makes communication more challenging as the security point of view. In previous sections, we mentioned the security threats and requirement for OppNets. Based on this discussion, we propose a security framework for OppNets in Fig. 2. Several other attempts made in this direction previously by Lilien et al. [6], Wu et al. [9], Poonguzharselvi et al. [10] are worth mentioning. Our framework has five modules.

Authentication is the first security module of our proposed framework. The function of authentication module is to prevent the unauthorized nodes from

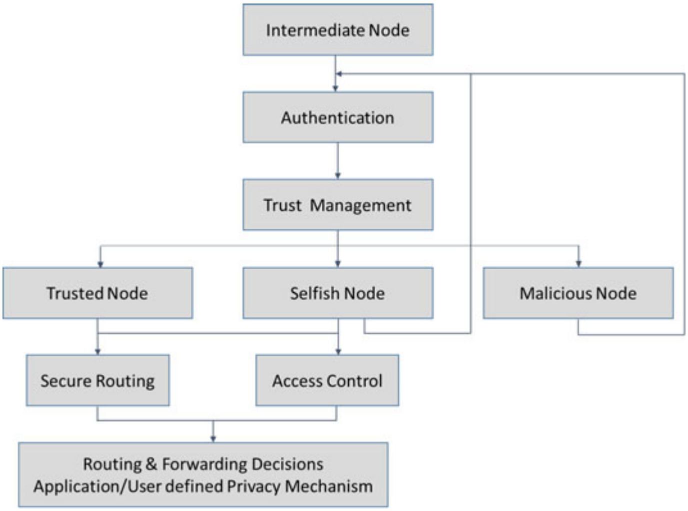


Fig. 2 Proposed security framework for OppNets

joining/accessing the network. Only the node verified by authentication module is allowed to join the OppNets. The second module is trust management. In trust management, the behavior of the node is continuously monitored and trust evaluation is done. On the basis of behavior analysis and trust evaluation nodes are categorized into three categories:

1. **Trusted Node**: These nodes maximize their contributions to the network community. These are the nodes with high trust level. These nodes are allowed to communicate to other nodes immediately.
2. **Selfish Node**: Selfish nodes are the nodes, which minimize their contributions, but maximize their individual gains by placing deceitful nodes into the network community or sometimes try to save their resources by dropping the packets. Such nodes may permit with their actions, but are continuously monitored by authentication and trust management modules.
3. **Malicious Node**: These nodes attack proper network operations, for example, spoiling the routing information, packet alteration, information mugging, etc., and do not consider their own gains. Such nodes are immediately stop to take part in any network activity and continuously being monitored.

Access control module is the third module of our proposed framework. This module puts restriction to the access of network resources and authorized the nodes to perform specific operations. Secure routing is fourth security module. In routing of OppNets, most of the current researches focus on context aware routing as OppNets are human centralized [11]. Currently no complete security scheme has been designed for OppNets routing protocols which ensure secure functioning of network. In our proposed framework this module is responsible for confidentiality, integrity threats. The last module describes about application/user defined privacy mechanism, to ensure the application and user-specific privacy, if any. As privacy is one of the main issues in OppNets due to the fact that many of the routing algorithms use the context information. This may be awkward to some users. Hence, this module allows nodes to show/hide information about them according to their wish.

6 Conclusion and Future Research Directions

As an emerging network paradigm, OppNets have a prominent vision in many applications area such as disaster and rescue networks, wildlife monitoring and tracking systems, social networking, providing connectivity in remote areas, and many more. Due to its *opportunistic* nature, OppNets are adaptive and can be used in different networking applications, but the security issues, mainly trust and privacy, defies OppNets to be widely used. Security is still an open and wide area in OppNets research. In this paper, we discussed about the security requirements for OppNets and point out how *PAINT* properties are important and needed to address for any security solutions for OppNets. Further, we discussed about the security threats in OppNets and finally proposed a security framework for OppNets.

References

1. Lilien, L., Gupta, A. & Yang, Z.: Opportunistic Networks for Emergency Applications and Their Standard Implementation Framework. *2007 IEEE Int. Perform. Comput. Commun. Conf.* (2007)
2. Sadler, C. M. & Martonosi, M.: Implementing Software on Resource-Constrained Mobile Sensors : Experiences with Impala and ZebraNet. *MobiSys '04 Proc. 2nd Int. Conf. Mob. Syst. Appl. Serv.* (2004). 256–269
3. Pentland, A., Fletcher, R. & Hasson, A.: DakNet: Rethinking Connectivity in Developing Nations. *Computer (Long. Beach. Calif).* 37, (2004). 78–83
4. PodNet - Mobile Distribution of User-Generated Content [PodNet Project]. at <http://www.podnet.ee.ethz.ch/>
5. Wang, W., Guo, F., Zheng, F., Tang, W. & Wang, J.: Research on Routing Protocols and Simulation Analysis for Opportunistic Networks. 10, (2015). 181–202
6. Lilien, L., Kamal, Z. H., Bhuse, V. & Gupta, A.: Opportunistic Networks : The Concept and Research Challenges in Privacy and Security. *Proc WSPWN* 140, (2006). 134–147
7. Chlamtac, I. & Lerner, A.: Fair Algorithms for Maximal Link Activation. *IEEE Trans. Commun.* 35, (1987). 739–746
8. Crowcroft, J., Hand, S., Mortier, R., Roscoe, T. & Warfield, A.: Plutarch: An Argument for Network Pluralism. *Appl. Technol. Archit. Protoc. Comput. Commun.* 33, (2003). 258–266
9. Wu, Y., Zhao, Y., Riguidel, M., Wang, G. & Yi, P.: Security and Trust Management in Opportunistic Networks : A Survey. *Secur. Commun. NETWORKS* 8, (2015). 1812–1827
10. Poonguzharselvi, B. & Vetriselvi, V.: Trust Framework for Data Forwarding in Opportunistic Networks Using Mobile Traces. 4, (2012). 115–126
11. Xi, C., Youliang, Guangsong, L. & Jianfeng, M.: Security in Opportunistic Networks. *Int. Conf. Ind. Control Electron. Eng.* (2012). 2006–2009

Replica-Based Efficient Data Accessibility Technique for Vehicular Ad Hoc Networks

Brij Bihari Dubey, Rajeev Kumar, Naveen Chauhan and Narottam Chand

Abstract Vehicular ad hoc networks are also one of the emerging fields which carry and forward data to deliver it to the destination. The proposed mechanism discusses scheme for replica augmentation when new replica is required, and replica abandon when any replica is disused from minimum fixed time period. Proposed scheme also performs arrangement of data into most relevant replica depending upon size of data stored, size of buffer left at each replica, deadline, and popularity of data. The proposed scheme also considers probability functions to fairly select most suitable replica for increasing accessibility.

Keywords VANETs · Scheduling · Road side unit · Service area · Vehicle-to-Vehicle (V2V) · Vehicle-to-Infrastructure (V2I)

1 Introduction

In the early days, safety was the key area of research in the field of VANETs. With the advancement in the technology few more areas have been evolved by researchers (like location aware information, traffic control, entertainment, etc.) and have emerged as the pivot areas of research where researchers are concentrating in recent years. In next generation wireless networks low cost high performance communication technologies are becoming epidemic due to its affordability. Drivers

B.B. Dubey (✉) · R. Kumar · N. Chauhan · N. Chand
Department of Computer Science and Engineering, National Institute of Technology, Hamirpur, India
e-mail: dubey.brijbihari@gmail.com

R. Kumar
e-mail: eminentpearl@gmail.com

N. Chauhan
e-mail: naveenchauhan.nith@gmail.com

N. Chand
e-mail: nar@nith.ac.in

P.K. Sa et al. (eds.), *Progress in Intelligent Computing Techniques: Theory, Practice, and Applications*, Advances in Intelligent Systems and Computing 719, DOI 10.1007/978-981-10-3376-6_51

and passengers can access services for which they have requested. The replication is the widely used technique which enhances data accessibility in highly mobile vehicular network.

There are two basic type of replication: one is replica of whole content collected at any node and other is replica of selected content from different nodes moving across road. The vehicular system has two types of replication mechanism: (a) System having replica such a way that Road Side Unit (RSU) is aware about replica locations. (b) System designed in such a way that RSU is not aware about replica locations. It has been proved that replica placement keeping RSU aware of it is NP-complete problem.

In the proposed work, authors have proposed a scheme to select most suitable data, and most suitable node to replicate data in such a way the vehicles can get required data with high probability with in limited range. The proposed technique takes care of requesting vehicles with packet level data transmission. Each vehicle selects data depending upon Encounter probability of two nodes, contact duration on each encounter. Data replication is very important and extremely challenging in distributed systems specially VANETs where vehicle changes their location and neighbors with alacrity. Replication not only increases accessibility in the network but also decreases load on individual RSU.

The remainder of the paper is organized as follows: Section 2 presents related work done by other authors in this field. Section 3 justifies the problem. Section 4 proposes solution and analyses proposed scheme. Section 5 draws some concluding remarks.

2 Related Works

In [1], Marco et al. discuss about algorithm to replicate data on nodes such that more data is replicated with the neighboring nodes. In this proposal authors does not deal with popularity of data. In [2], Fiore et al. propose content replication algorithm, called Hamlet, which keeps track of neighbor vehicles content to target different content than of neighbor vehicles so that node carry more data in surroundings. In [3], Silva et al. propose destination based replication mechanism which selects specific vehicle for replication depending upon specific criteria. In [4], Gossa et al. propose a data replication mechanism which takes vehicle mobility into account and predicts its motion. In [5], Li et al. propose contact aware data replication mechanism which is applied to replicate near roadside content in Road Side Units (RSUs). The content selected for replication are collected by analyzing real mobility traces based on the analysis of real mobility traces in terms of contact frequency and contact duration. In [6], Bruno et al. propose a technique which takes popularity into consideration this optimization technique ameliorates content accessibility in the network.

3 Problem Statement

In the vehicular network, due to high topology change, it is challenging to establish end-to-end path and due to mobility connectivity exists for limited time period. In the vehicular network, node requests for their required data items and all other nodes in the network including RSU aims to serve the request. Since the nodes have limited storage space each node cannot keep all the information that is demanded by vehicles. On the other hand, if all information is managed to store at each and every node then network performance degrades severely due to high maintenance and update cost. To solve this problem, whenever any node request for specific data item, it searches required data with itself. If node itself is not carrying that data item, it searches in the neighbors, if it does not find this data in the neighbor within limited time, it forwards that request to the RSU. Each RSU typically receives large number of such requests and processes them by applying scheduling algorithms by taking few parameters into account. Therefore, the objective is set to satisfy all the vehicles requesting the data having fresh copies within the limited time frame.

3.1 Problem Formulation

There are set of vehicles V_r requesting for data items and set of vehicles V_c containing required data that has been requested by $v_r \in V_r$. When vehicle $v_c \in V_c$ is at h hop away from the reach of the nearest requesting vehicle then it starts replication process by selecting each vehicle encountering. When a Vehicle $v_k^L \in V_L$ is encountered to $v_c \in V_c,$ it can receive data from v_c to satisfy all requesting vehicles. The objective function is formulated to distribute the required data in the neighbor such each requesting vehicle get at least one copy of requested data within at most h hop distance. The objective function is described using $M_q^{v_r}(t_1)$ which is defined as data items requested by vehicle v_r at time t_1 containing unique data identity (M_q). So, to satisfy requesting vehicles in the vehicular network the objective function is set as max(X) *s. t.* requested message $\{M_q : q = 1, 2, \ldots, m\}$, which have been requested by vehicles $v_i^r \in V_r$ of the target region. There are N_i^q destinations that have requested message M_q. If message M_q has n_d copies, it has to be distributed in the target area in such a way that each requesting vehicle finds its requested data within h hop distance. The vehicles $v_j^c \in V_c$ contain reply message has $\{M_q^{'v_r} : q = 1, 2, \ldots, m\}$ in such a way that $\max\left(\bigcup_{v_i \in V_L} M_q^{'v_i}(t_1)\right)$. Subject to $\bigcup_{v_j \in V_r} M_q^{v_j}(t_1) \subseteq \bigcup_{v_i \in V_L} M_q^{'v_i}(t_1)$ and $\left|v_i^r v_j^L\right|^{new} \leq D(h)$. Where, $D(h)$ is the distance between requesting vehicle and vehicle this replicating data in its neighbor, and $\left|v_i^r v_j^L\right|^{new}$ is the current distance between vehicle carrying data and requesting vehicle (as shown in Fig. 1). The condition-mentioned above shows that all the required data is replicated in the vehicles near the target region and is distributed to

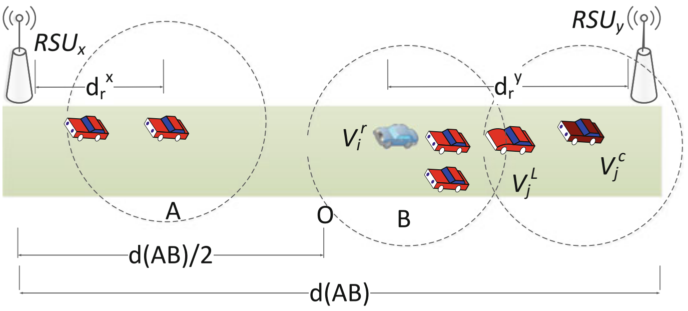


Fig. 1 Scenario of vehicles requesting for data

the vehicles of the region in such a way that the requested data is maximum h hops away.

4 Proposed Solution

The proposed solution follows following procedure to deal with requests send by various requesting nodes.

4.1 Proposed Mechanism

In the paper, authors propose a mechanism which selects data required by $v_i^r \in V_r$ to forward first. When data carrying vehicle v_j^c reaches in the h hop communication range of $v_i^r \in V_r$. The first vehicle found closest to h hop distance is given priority to get served. Since, the vehicles requests for several items of different data size, the number of data items requested is also different. For data M_q if $M_q \geq \sigma * \tau$, this implies that total required data items cannot be transferred to a single vehicle and vehicle waits for one more encounter with any other vehicle. If after receiving $\sigma * \tau$, data items vehicle disappears and later again encounters and successfully establishes communication then this encounter is treated as fresh encounter and the same vehicle can contend for getting remaining data items. If any vehicle interacts then the remaining items are transmitted to that vehicle. The number of data items transmitted to this vehicle, again, depends upon contact duration. In this scenario whenever any vehicle encounters, the data items are forwarded to it and the remaining data items are forwarded to vehicle observed during next encounter. This process is continued till all the data items are forwarded. Once all the data items are transmitted to selected set of vehicles request v_L, they are supposed to keep a copy of data as replica and also, now it is checked that the distance from nearest requesting vehicle (for data d) is not more than h hop (The discussion of finding number of hops is discussed later). If total number of hops is less than h, the same process is repeated again for all next encounters till total number of hop reaches to h. If the total number of hop reaches to h after few cycles of data forwarding, the vehicle starts forwarding other requested data. When all the data of requesting vehicle v_i^r is completed, the data required by other vehicle v_j^r is forwarded using same mechanism till all the data items are transmitted. This process is repeated till vehicle v_k^c containing required data transmits to the vehicles encountered and/or reaches out of the h hop range of requesting vehicle. If total number of hop reaches to h before completing a single cycle of forwarding all required data items then the size of data transmitted on each encounter is increased by factor ς. Again if $p_s \leq \sigma * \tau$, the vehicle encountered receives all p_s data items. The vehicle v_i^r utilizes remaining time in transmitting other data that are requested by same or other vehicle depending upon availability.

If during encounter with any vehicles communication establishes between them, the max data that can be transmitted is given by x. To reduce the complexity of calculation, it is assumed that average contact duration between two vehicles is given by τ. The maximum number of data items that can be transferred during contact period is given by $x = \sigma * \tau$. Since, more than one vehicle can arrive in the communication range of vehicle v_k^c, but data carrying vehicle can communicate with only one vehicle at a time (see assumption 2). So, if p data items are required to reply any query q. The time taken to contact one of these vehicle is given by $1/\rho$. Probability of encountering any vehicle v_L in above time is given by time $(1/\rho)(1 - e^{-\rho})$. Assume that probability of staying any vehicle for τ period of time is $P(A_{\mathrm{v_r}}^{\tau})$. Probability of transmitting $\sigma * \tau$ data items that can be transmitted to vehicle encountered is represented by

$$P(E) = \frac{\sigma * \tau}{\rho}(1 - e^{-\rho})P(A_{\mathrm{v_r}}^{\tau}). \tag{1}$$

4.2 Node Selection Policy for Replica Placement

Since, in VANETs several nodes arrive simultaneously in communication of each other to receive as well as forward or replicate data. Also, typically in dense networks several vehicles encounters to each other in wee hour and in that short time it is challenging to communicate with each other. In this type of scenario, it is a challenge to select vehicle for replicating copy of data. For this purpose, polling is done among all encountered node to select find node to replicate selected data. If two nodes have the same priority to get selected for placing replica, the node having

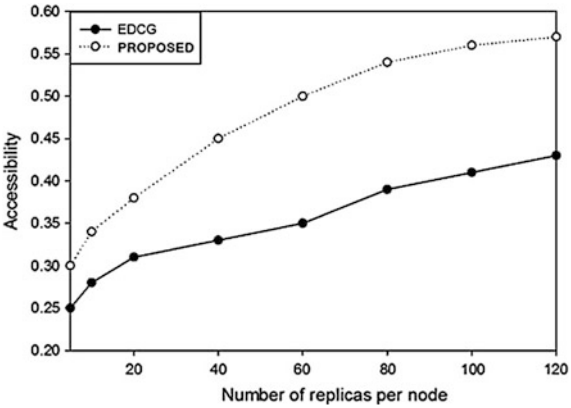


Fig. 2 The effect of the number of replicas per node on data accessibility

relative speed closure to average speed is selected as replica. The proposed mechanism is compared with EDCG scheme [7] and found that proposed schemes results are satisfactory (as shown in Fig. 2).

5 Conclusion

Vehicular network has envisaged researchers to strengthen intelligent transportation system. In the proposed protocol, we have discussed replication mechanism which allows vehicles to keep replica of forwarded data items. Since, replica placement problem is NP-hard and it is unrealistic to propose exact solution for it. The proposed solution increases accessibility in the VANETs and it is good approximation of replica selection. In future, we would like to optimize proposed solution and work on finding methods to achieve closure approximation of number of nodes to replicate data.

References

1. Fiore, M., Casetti, C. & Chiasserini, C.-F. Caching strategies based on information density estimation in wireless ad hoc networks. *Veh. Technol. IEEE Trans.* 60, 2194–2208 (2011).
2. Trullols-Cruces, O., Fiore, M. & Barcelo-Ordinas, J. M. Cooperative download in vehicular environments. *Mob. Comput. IEEE Trans.* 11, 663–678 (2012).
3. Silva, F. A. *et al.* ODCRep: Origin–Destination-Based Content Replication for Vehicular Networks. *Veh. Technol. IEEE Trans.* 64, 5563–5574 (2015).
4. Gossa, J., Janecek, A. G., Hummel, K. A., Gansterer, W. N. & Pierson, J.-M. Proactive replica placement using mobility prediction. in *Mobile Data Management Workshops, 2008. MDMW 2008. Ninth International Conference on* 182–189 (2008).
5. Li, Y., Jin, D., Hui, P. & Chen, S. Contact-aware data replication in roadside unit aided vehicular delay tolerant networks. (2015).

6. Bruno, F., Cesana, M., Gerla, M., Mauri, G. & Verticale, G. Optimal content placement in icn vehicular networks. in *Network of the Future (NOF), 2014 International Conference and Workshop on the* 1–5 (2014).
7. Hara, T. Replica allocation methods in ad hoc networks with data update. *Mob. Networks Appl.* 8, 343–354 (2003).

Reinforcement Based Optimal Routing Algorithm for Multiple Sink Based Wireless Sensor Networks

Suraj Sharma, Azad Kumar Patel, Ratijit Mitra and Reeti Jauhari

Abstract Routing in wireless sensor networks has been a recent area of research because of its increased use in diverse application environments. Sensor nodes are energy constrained which possess a formidable challenge in designing efficient routing algorithms for them. Most of the scenarios where sensor networks are used such as battle field surveillance, health monitoring are delay sensitive in nature. To mitigate these problems, we have proposed two routing algorithms in this paper: one based on multiple static sink based scenario and the other based on multiple mobile sink based scenario. Both of these protocols use reinforcement learning methodology in order to solve the routing problem in an intelligent and efficient way.

Keywords Sensor · Reinforcement · Routing · Sink

1 Introduction

Wireless sensor networks consist of tiny nodes which can sense, compute, and communicate in wireless medium. The data is sensed and collected by these nodes and are routed to the sink nodes. Wireless sensor network have become an important area of research because of their use in diverse application areas like battle field surveillance, environmental monitoring, disaster relief operations, home security systems, forest fire monitoring, animal habitat monitoring, nuclear firm monitoring, and many

S. Sharma (✉) · A.K. Patel · R. Mitra · R. Jauhari
Department of Computer Science and Engineering, International Institute
of Information Technology, Bhubaneswar, India
e-mail: suraj@iiit-bh.ac.in

A.K. Patel
e-mail: patel.azad55@gmail.com

R. Mitra
e-mail: ratijit.mitra@gmail.com

R. Jauhari
e-mail: johri0921@gmail.com

P.K. Sa et al. (eds.), *Progress in Intelligent Computing Techniques: Theory, Practice, and Applications*, Advances in Intelligent Systems
and Computing 719, DOI 10.1007/978-981-10-3376-6_52

more [1, 2]. They are generally operated by battery power which exhausts quickly and is difficult to replace. Therefore, energy needs to be optimally used to enhance network lifetime and overall functionality of the network. Moreover, the application areas where they are used are delay sensitive in nature so end to end delay too, needs to be minimized. To mitigate the above problems two intelligent routing algorithms based on multiple static and multiple mobile sink have been proposed in this paper.

The rest of paper is organized as follows: Sect. 2 states the related work, Sect. 3 describes the network assumptions and data structures used, Sect. 4 describes the first proposal that is multiple static sink-based routing protocol, Sect. 5 describes the second proposal that is multiple mobile sink-based routing protocol, Sect. 6 gives a theoretical analysis to the proposed works, and Sect. 7 concludes the paper.

2 Related Works

A number of routing protocols with varying network architectures have been proposed in literature. Some of them are Grid based, Hierarchical, location based, flat, and hybrid (consists of a combination of one or two of the above routing architectures) [7]. In TTDD proposed by H. Luo et al. in [3] complexity in the construction of grid structure is the major hindrance. In HExDD proposed by A.T. Erman et al. in [6] center nodes become the major hot spot thereby decreasing network lifetime. In HCDD proposed by C.J. Lin et al. in [4] Load balancing is not implemented. In GBEER proposed by K. Kweon et al. in [5] the nodes in the quorum can become an overhead and decrease network lifetime. In VGDD proposed by A.W. Khan et al. in [8] end-to-end delay maybe more as the partial path is not likely to be optimal. In FTIEE proposed by F. Kiani et al. in [9] mitigate above problems in an intelligent and efficient way.

3 Network Assumptions and Data Structures

The deployment area is square and uniform in topology with all the nodes location aware. Sensors and sink nodes are deployed in a random fashion. The topology is represented by fully connected graph as shown in Figs. 1 and 2. The data structures and terminologies used are enumerated below:

1. GN: Gateway nodes formed after routing and learning process for each of the subregions. They are used to route data to the respective sink as shown in Fig. 5.
2. $Q[N_i]$: Q-value of a node according to which routing decision is taken.
3. n: Number of sensor nodes in the deployed region.
4. z: Number of neighbors of node N_i where $1 \leq i \leq n$.
5. m: Number of sensor nodes in the sub-region space.

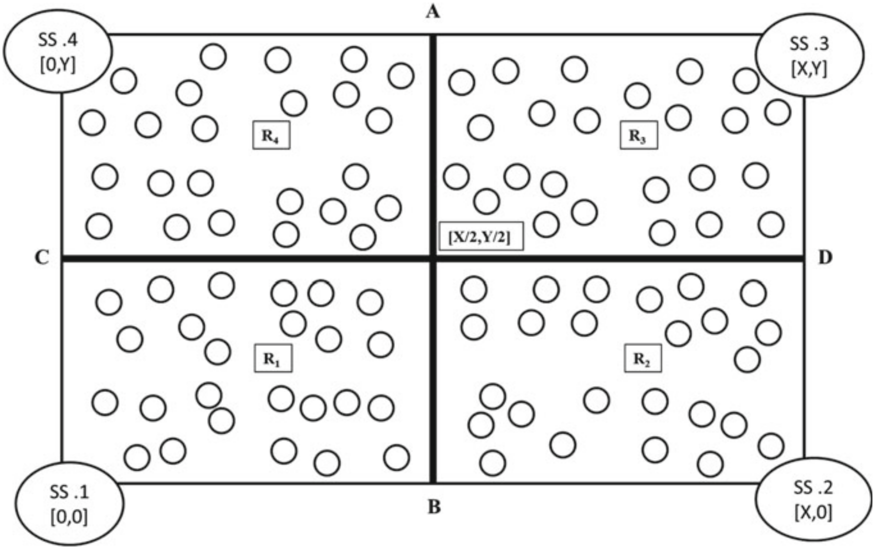


Fig. 1 Static sink allocation

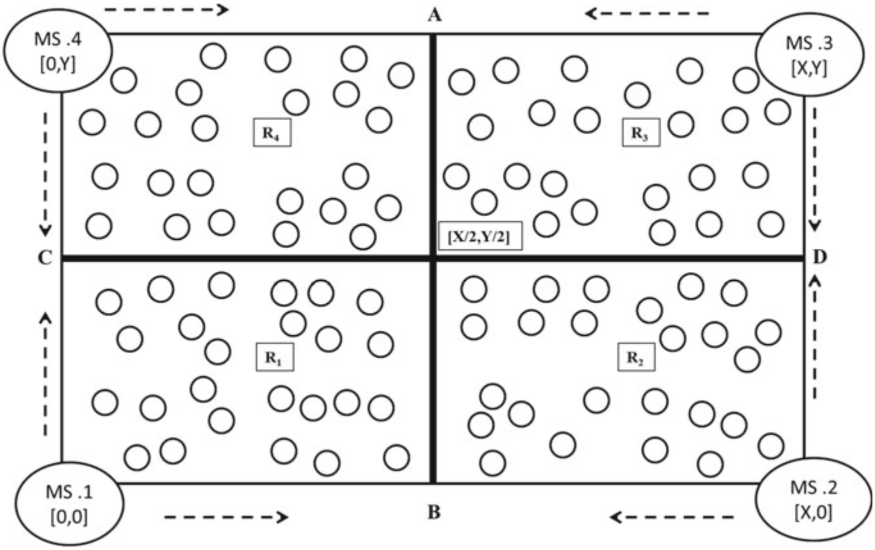


Fig. 2 Mobile sink allocation

$$m = (\lfloor n/4 \rfloor) \tag{1}$$

6. SS.i: Denotes static sink nodes represented by ellipses.
7. MS.i: Denotes mobile sink nodes represented by ellipses.

8. SINK: Denotes sink nodes (may be static or mobile) represented by ellipses as shown in Figs. 1 and 2.
9. (S):Denotes source node.
10. $B[N_i]$: Denotes battery value of the sensor node where $1 \leq i \leq n$.
11. β: Denotes the threshold value for Q-value calculation.
12. $SR\_ID[N_i]$: Region to which the sensor nodes or sink nodes belong respectively, where $1 \leq i \leq n$.
13. $LOC\_ID[N_i]$: Denotes the coordinate positions (u_i, v_i) of the sensor and the sink nodes, where $1 \leq i \leq n$.
14. $C[N_i]$: Denotes the cost value of each sensor node as calculated by Algorithm 2, where $1 \leq i <= n$.
15. $d[N_i]$: Euclidean-Distance of sensor nodes from their respective sinks, where $1 \leq i \leq n$.
16. V: Velocity of the sink moving in the direction as shown in Fig. 2.
17. D: Distance traveled by the sink moving in direction as shown in Fig. 2. Here the distance is the length of the one side of the subregion as depicted in Eq. 2.

$$D = X/2 \tag{2}$$

18. M_T: Time taken by the sink to cover one round of motion during the motion phase as depicted in Eq. 3.

$$M\_T = D/V \tag{3}$$

19. S_T: Sojourn time of the sink nodes as depicted in Eq. 4.

$$S\_T = \frac{\sqrt{(X/2)^2 + (Y/2)^2}}{V} \tag{4}$$

20. $LOC\_ID[N_i]$: Denotes the coordinate positions (u_i, v_i) of the sensor and the sink nodes, where $1 \leq i \leq n$.
21. NBR_SRCH: Probe packets sent by the sensor nodes for neighbor detection. Contains $NBR\_TAB[N_i]$ where i is the number of neighbor nodes.
22. NBR_SENT: a flag that is set to true when the NBR_SRCH (neighbor probe packet) has been sent.
23. $BRDCST\_PCKT$: This packet broadcasts the changed table of node to its other neighbors.
24. NBR_SIG: A type of beacon send along with $BRDCST\_PCKT$.
25. $NBR\_TAB[N_i]$: The neighbor list maintained at each of the sensor nodes contains battery value $(B[N_i])$, $(LOC\_ID[N_i])$, $(SR\_ID[N_i])$, $(C[N_i])$ where $1 \leq i \leq z$.

4 Proposed Approach I: Multiple Static Sink

An intelligent routing protocol based on reinforcement learning methodology, multiple static sink-based intelligent routing protocol (MSSIR) is described. It consists of three phases subregion allocation phase, neighborhood detection phase, routing phase. In the first phase, the entire region is divided into four equal sized quadrants. Each of the sensor nodes and the sink nodes are allocated to the specific quadrants by using subregion allocation algorithm as shown in Algorithm 1.

Then as all the nodes in the static sink-based frame work are acquainted with their neighbors by neighborhood detection algorithm as discussed in Algorithm 5, cost calculation as described in Algorithm 2 is executed and cost value for each node is computed. At the end of this phase, all the nodes know their respective costs.

Then the routing phase starts. In this phase, first of all the nodes receive the cost of their neighbors. After all the nodes are acquainted with their neighbor cost and one of the nodes has some data to send, it calculates the Q-values as discussed in Algorithm 3 and forwards the data along the path that has the maximum Q-value of its nodes as described by Algorithm 4. The figures for the various phases are depicted in Figs. 3, 4, and 5.

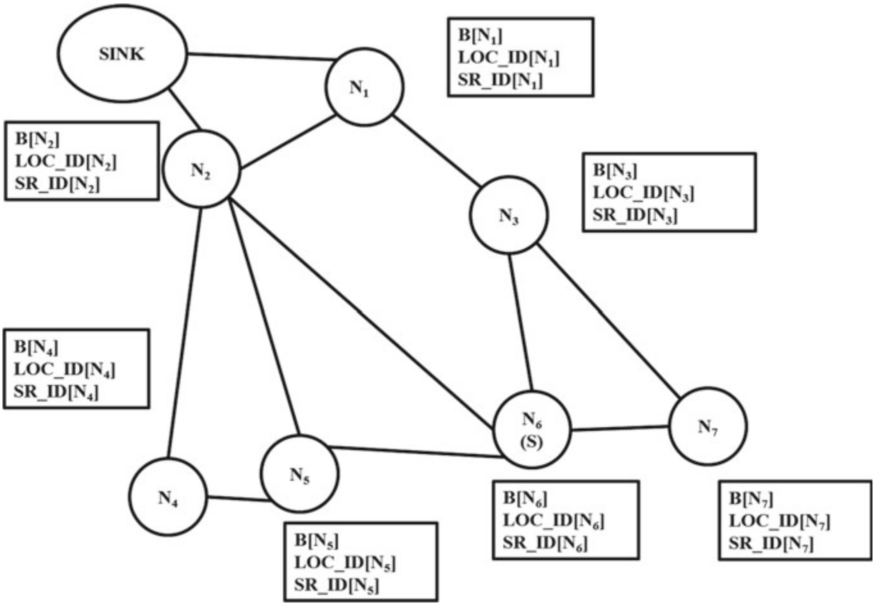


Fig. 3 Network topology before neighborhood detection

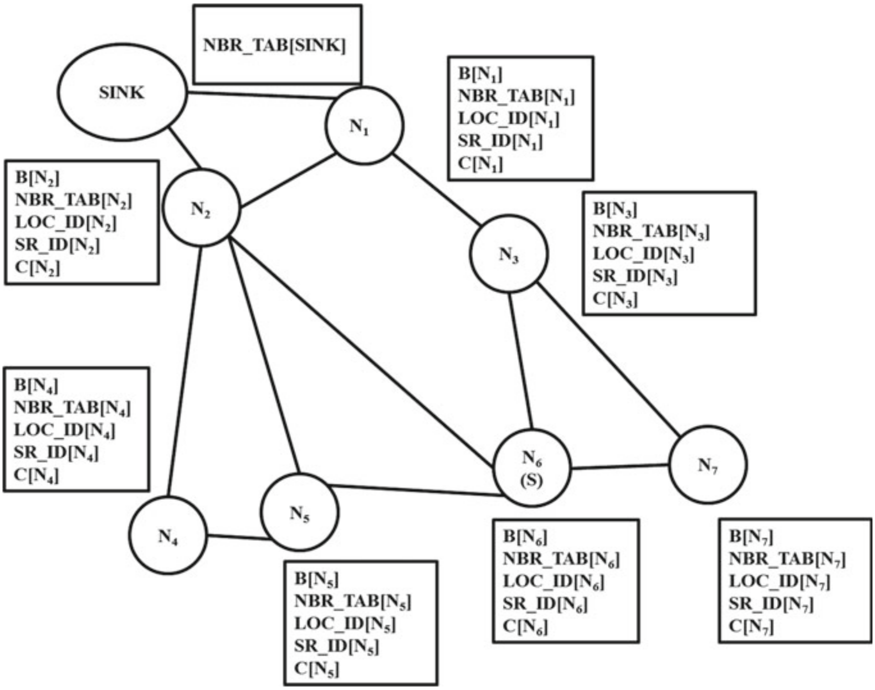


Fig. 4 Network topology after neighborhood detection

Algorithm 1 Sub-Region Allocation

```
if ($u \geq 0$ and $u \leq X/2$ and $v \geq 0$ and $v \leq Y/2$) then
    ($SR\_ID[N_i] = 1$)                                  ▷ The first quadrant of the deployed area
end if
if ($u \geq X/2$ and $u \leq X$ and $v \geq 0$ and $v \leq Y/2$) then
    ($SR\_ID[N_i] = 2$)                                  ▷ The second quadrant of the deployed area
end if
if ($u \geq X/2$ and $u \leq X$ and $v \geq Y/2$ and $v \leq Y$) then
    ($SR\_ID[N_i] = 3$)                                  ▷ The third quadrant of the deployed area
end if
if ($u \geq 0$ and $u \leq X/2$ and $v \geq Y/2$ and $v \leq Y$) then
    ($SR\_ID[N_i] = 4$)                                  ▷ The fourth quadrant of the deployed area
end if
```

Algorithm 2 Cost Computation at Each Node

```
if ($SR\_ID[N_i] == 1 || SR\_ID[N_i] == 2 || SR\_ID[N_i] == 3 || SR\_ID[N_i] == 4$ ) then
    For all nodes ($C_{sink}[i] = m$) For each of the neighbors of sink[i] for instance x $C_x$ is computed as:

        $$C_x = C_{sink}[i] * (1 - (1/(m-1)) + (1/d_{sink[i],x}) \quad (5)$$

      sink[i] sends $C_x$ to node x
      Each neighbors such as p receives $C_x$ for itself and then compute new $C_i$ for each of its neighbors as follows:

        $$C_i = C_p * (1 - (1/(m-1)) + (1/d_{p,i}) \quad (6)$$

  Then it sends $C_i$ to its neighbors and the process is repeated.
end if
```

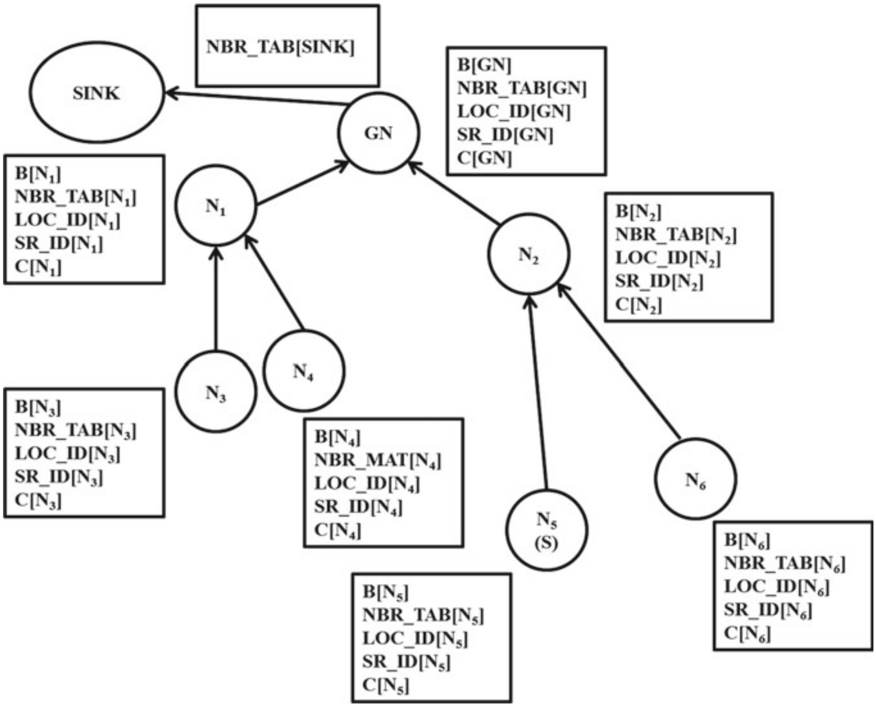


Fig. 5 Network topology after routing

Algorithm 3 Q-value Computation at Each Node

```
if ($SR\_ID[N_i] == 1$ || $SR\_ID[N_i] == 2$ || $SR\_ID[N_i] == 3$ || $SR\_ID[N_i] == 4$ ) then
    For the neighbors of i for instance p $Q[N_i]$ is computed as:
    if ($B_i \geq \beta$) then
        $Q[N_i] = 1/3 * B_p + 2/3 * C_p$    (7)
    else
        $Q[N_i] = 2/3 * B_p + 1/3 * C_p$    (8)
    end if
end if
```

Algorithm 4 Routing in the Static Sink-Based Topology

```
if ($SR\_ID[N_i] == 1$ || $SR\_ID[N_i] == 2$ || $SR\_ID[N_i] == 3$ || $SR\_ID[N_i] == 4$ ) then
    1. Each node in each sub region sends a message to its entire neighbor to receive their costs.
    2. When the source node say s has a packet to send within its sub-region. And its neighbors are z[i]
    if ($SR\_ID[s] == SR\_ID(z[i])$ and $cost(z[i]) > cost[s]$) then
        (Compute Q-value of z[i] and send the packet to the one having highest Q-value.)
    else(Send to any of the neighbors randomly)
    end if
end if
```

5 Proposed Approach II: Multiple Mobile Sink

Here the second variant Multiple Mobile Sink-based Intelligent Routing Protocol (MMSIR) is described. The general structure of the sensor network topology with multiple mobile sinks is depicted as in Fig. 2. Each of the mobile sinks are allocated to the different subregions as in case of static sink. Then the sink trajectory is decided as depicted in Fig. 2. The mobility of the sink makes the sink to acquire two states during the entire protocol operation. One is motion state and the other is sojourn state. The sojourn state is again divided into two time periods. In the motion state, the sink nodes move along the predetermined trajectory as shown in Fig. 2. In the sojourn state, during the first time period neighborhood detection is done and in the second time period routing is done.

The entire protocol consists of three phases such as subregion allocation phase, neighborhood detection phase, and routing phase. The subregion allocation phase follows the similar procedures as for static sink case.

Then comes the neighborhood detection phase. In this as the sink is mobile so the neighborhood of the sink changes frequently [10]. In this case we need to flood the neighborhood matrix in the first half of the sojourn time-period as described in Algorithm 5.

Routing phase generally begins in the second half of the sojourn time period of the mobile sink. In this the nodes first of all receive the cost of their neighbors. Whenever the node has some data to send it calculates the Q-values and forwards the data along the path that has the maximum Q-value. The figures for the various phases are depicted in Figs. 3, 4, and 5.

Algorithm 5 Neighbor-hood Detection in the Mobile Sink-Based Topology

```
if ($SR\_ID[N_i] == 1 \,||\, SR\_ID[N_i] == 2 \,||\, SR\_ID[N_i] == 3 \,||\, SR\_ID[N_i] == 4$ ) then
    Suppose node a receives following packet from node b
$NBR\_SRCH\ :< b[NBR\_SRCH, NBR\_TAB(b)] >$
    if ($LOC\_ID(b) \notin NBR\_TAB(a)$) then
        $NBR\_TAB[a] = NBR\_TAB(b) \cup NBR\_TAB[a]$
        if ($NBR\_SENT(a)) == false$) then
            $(NBR\_SENT(a)) == true)$ $BRDCST\_PCKT(NBR\_SIG, NBR\_TAB(a))$        ▷ Broadcast neighbor probe
packet
        elseDrop the Packet
        end if
    elseDrop The Packet
    end if
end if
```

Algorithm 6 Routing in the Mobile-Sink based Topology

```
if ($SR\_ID[N_i] == 1 \,||\, SR\_ID[N_i] == 2 \,||\, SR\_ID[N_i] == 3 \,||\, SR\_ID[N_i] == 4$ ) then
    For the first half of the sojourn time find the neighborhood by neighborhood detection algorithm
    For the second half of the sojourn time
    Each node in each sub region sends a message to its entire neighbor to receive their costs
    When the source node say s has a packet to send within its sub-region. And its neighbors are z[i]
    if ($SR\_ID[s] == SR\_ID(z[i])$ and $cost(z[i] > cost[s])$) then
        (Compute Q-value of z[i] and send the packet to the one having highest Q-value.)
    else(Send to any of the neighbors randomly)
    end if
end if
```

6 Theoretical Analysis

In this paper, we have proposed two variants of an intelligent routing protocol in our work. One consisting of a multiple static sink approach and the other considering a multiple mobile sink approach. The first approach helps in reduction of end-to-end delay as compared to the previous work consisting of single sink.

The second approach consisting of multiple mobile sinks helps in resolving the energy-hole problem thereby enhancing network lifetime. Also, end-to-end delay is minimized compared to the previous work consisting of single sink.

We have also eliminated the step of grid-construction from the previous work which was an unnecessary overhead, thereby reducing the control packet length and hence congestion in the network.

7 Conclusion

In this paper, we propose two routing protocols. The first work proposes multiple static sink and helps in reduction of end-to-end delay in the network. The second one uses a multiple mobile sink-based solution and helps improve network lifetime by eliminating the energy-hole problem along with giving the benefits as in multiple static sink case. We have proposed both the routing protocols by using reinforcement-based learning approach. This approach uses the Q-value technique to optimally resolve the above-sated problems.

References

1. K. Akkaya and M. Younis, *A Survey On Routing Protocols For Wireless Sensor Networks*, Ad Hoc Networks, Volume 3, pp 325–349, November, 2005.
2. Jamal N. Al-Karaki and Ahmed E. Kamal, *Routing Techniques in Wireless Sensor Networks: A Survey*, IEEE Wireless Communications, Volume 11, Issue 6, pp 6–28, December, 2004.

3. H. Luo, F. Ye, J. Cheng, S. Lu and L. Zhang, *TTDD: Two Tier Data Dissemination in Large Scale Wireless Sensor Networks Sensor Networks*, Wireless Networks, Volume 11, Issue 1, pp 161–175, January, 2005.
4. C. J. Lin, P .L. Chou, C. F. Chou, S. Lu and L. Zhang, *HCDD: Hierarchical Cluster-based Data Dissemination In Wireless Sensor Networks with Mobile Sink*, In the proceedings of International Wireless Communications and Mobile Computing Conference (IWCMC'06), pp 1189–1194, July, 2006.
5. K. Kweon, H. Ghim, J. Hong and H. Yoon, *Grid-Based Energy-Efficient Routing from Multiple Sources to Multiple Mobile Sinks in Wireless Sensor Network*, In the Proceedings of the 4th IEEE International Conference on Wireless pervasive computing, pp 185–189, 2009.
6. A. T. Erman, A. Dilo and P. Having, *A Fault-Tolerant Data Dissemination Based on Honeycomb Architecture For Mobile Multi-Sink Wireless Sensor Networks*, In The Proceedings Of IEEE Sixth International Conference on Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP'10), pp 97–103, 2010.
7. C. Tunca, S. Isik, M. Y. Donmez and C. Ersoy, *Distributed Mobile Sink Routing For Wireless Sensor Networks: A Survey*, Communications Surveys and Tutorials, IEEE Volume 16, Issue 2, pp 877–897, October, 2013.
8. A. W. Khan, A. H. Abdullah, M. A. Razzaque, J. I. Bangash and A. Altameem, *VGDD: A Virtual Grid Based Data Dissemination scheme for Wireless Sensor Networks with Mobile Sink*, Hindawi Publishing Corporation International Journal of Distributed Sensor Networks, Volume 2015, pp 1–16, December, 2014.
9. F. Kiani, E. Amiri, M. Zamani, T. Khodadadi and A. A. Manaf, *Efficient Intelligent Energy Routing Protocol in Wireless Sensor Networks*, Hindawi Publishing Corporation International Journal of Distributed Sensor Networks, volume 2014, pp 1–13, August, 2014.
10. S. Sharma and S. K. Jena, *Cluster based Multipath Routing protocol for Wireless Sensor Networks*, Hindawi Publishing Corporation International Journal of Distributed Sensor Networks, volume 2014, pp 1–13, August, 2014.

Study and Impact of Relay Selection Schemes on Performance of an IEEE 802.16j Mobile Multihop Relay (MMR) WiMAX Network

Chaudhuri Manoj Kumar Swain and Susmita Das

Abstract This paper investigates two efficient relay selection algorithms on the performance of an IEEE 802.16j MMR WiMAX network. The relay selection algorithms considered here are max-min relay selection and harmonic mean of SNR relay selection. The WiMAX transmitter comprises of OFDMA transmission technique and the receiver with maximum likelihood (ML) detection method is adopted for reliable detection of transmitted bits. The relays used here are the Amplify-Forward (AF) relay and Decode-Forward (DF) relay. Both selection combining (SC) and maximal ratio combining (MRC) techniques are applied and the performance metrics are symbol error rate (SER) and channel capacity. The analysis of the network is performed using the MATLAB software. It is observed that the DF relay along with MRC combination technique outperforms the other relay-diversity combining techniques in terms of providing improved SER and channel capacity. Again in providing less SER, the harmonic mean of SNR relay selection algorithm is superior in comparison to max-min relay selection algorithm. Further, the max-min based relay selection algorithm provides better channel capacity about the harmonic mean of SNR relay selection algorithm for the considered MMR WiMAX network.

Keywords WiMAX · OFDMA · AF · DF · SC · MRC

1 Introduction

For improving the throughput, coverage and system capacity in a broadband communication system, wireless relaying plays an important role [1]. IEEE 802.16j is a promising wireless access technology for broadband data service. It is designed

C.M.K. Swain (✉) · S. Das
National Institute of Technology, Rourkela 769008, India
e-mail: mnj.4444@gmail.com

S. Das
e-mail: susmita090991@gmail.com

P.K. Sa et al. (eds.), *Progress in Intelligent Computing Techniques: Theory, Practice, and Applications*, Advances in Intelligent Systems and Computing 719, DOI 10.1007/978-981-10-3376-6_53

to deliver high data rate with network coverage of 15 km for fixed users and 5 km for mobile users. To further enhance capacity, data rate for mobile stations at the cell edge, cooperative relaying technology with WiMAX is suggested in [2]. In cooperative communication, relay nodes are placed as intermediate nodes which help in forwarding the data packets from transmitter to the receiver with diversity gain. Although the relay nodes can increase system performance, use of single relay station suffers from fading effects. So to combat the effect of fading, WiMAX can use distributed space time code in which multiple relays are deployed and coordinate with each other to provide higher spatial diversity gain [3]. Selection of optimal relay station with the aim to maximize system capacity in IEEE 802.16j is suggested in [4]. In [5] the path metrics like available link bandwidth, Signal to noise ratio and hop count are used by each relay station in the 802.16j based MMR network for selecting the optimal path. Relay selection process holds an important role in acquiring high diversity order and providing optimum end to end channel capacity is illustrated in [6]. In [7] max-min based relay selection scheme is analyzed considering Rayleigh fading channel for both AF and DF relay protocol. Authors in [8], explore two algorithms with the aim of selecting a path for mobile nodes and for newly entering relays so as to maintain proper QOS requirements, link quality, and bandwidth availability. In [9], the authors propose two relay selection schemes named max-min cooperative relay selection scheme and joint cooperative relay-path selection scheme with the objective to achieve highest diversity gain and system throughput respectively. The improvement in FER performance and throughput using relay selection and power allocation algorithms are demonstrated in [10].

The contributions of the paper are as follows. Study and impact of relay selection algorithms using AF and DF relays with diversity combining techniques are demonstrated for the considered WiMAX network. Both the performance measures like SER and channel capacity are analyzed to prove the efficacy of relay selection schemes.

The rest of the paper is organized as follows. In Sect. 2, a brief discussion of the system model and its operation is provided. In Sect. 3, analytical description of the considered communication network is presented. Section 4 describes the two relay selection algorithms and the impact of these schemes on the performance metrics are analyzed in Sect. 5. Finally, the conclusion and future work are provided in Sect. 6.

2 System Model

In this paper, a point-to-point multi-relay based WiMAX network is considered which consists of a WiMAX transmitter, a WiMAX receiver and four number of relays. All the four relays are placed at an equal distance from the transmitter and the receiver. Relay link and access link represents the path between the transmitter-relay and relay-receiver respectively. The communication through

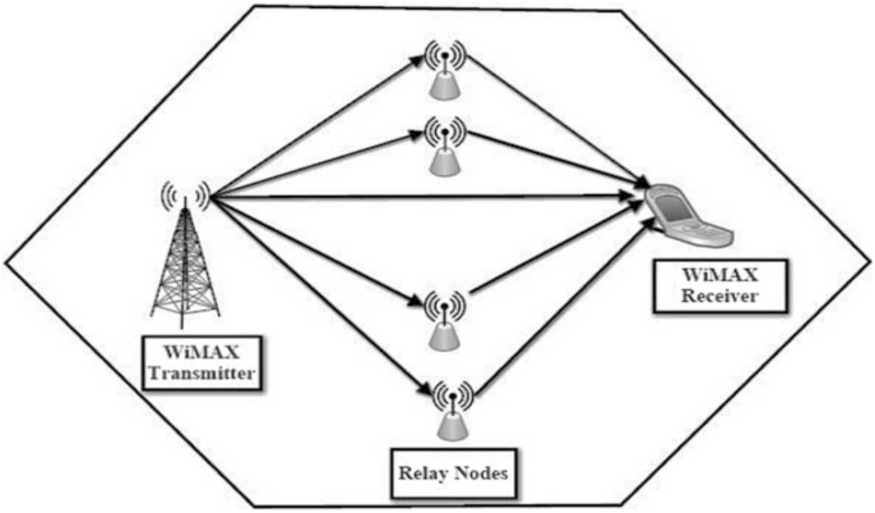


Fig. 1 System model of point to point MMR WiMAX network

relays is assumed as half-duplex and the information from the transmitter is propagated to the receiver in two phases, i.e., broadcast phase and cooperate phase. In the broadcast phase, the information is transmitted from the transmitter to the relays and in the cooperate phase, the information is delivered from the relay to the receiver. Figure 1 represents the system model of the multi-relay WiMAX network. The channel responses in both the links follow Rayleigh distribution. The noise contaminated with the information propagating in both the links is taken as Additive White Gaussian noise (AWGN) with zero mean and unit variance. In the relay node, two types of relays are taken for analysis purpose, AF and DF relay. At the receiver, diversity combining techniques, i.e., SC and MRC are used for combining the signals coming from the relays and the transmitter into the receiver. An optimum receiver named maximum likelihood (ML) receiver is used at the receiver for detecting the received signal in such a way to maximize the signal to noise ratio (SNR).

3 Analytical Description of the System Model

The output signal from the transmitter can be expressed as

$$x(n) = 1/\sqrt{N} \sum\nolimits_{n=0}^{N-1} X(k)\, e^{\frac{j2\pi nk}{N}}, \tag{1}$$

where '*N*' corresponds to the number of subcarriers used in the OFDM transmission technique, $X[k]$ stands for the symbol to be transmitted on the '*k*th' subcarrier.

For direct transmission, the received signal on the '*k*th' subcarrier at the receiver from the transmitter is expressed as

$$Y_{s,\ d}(k) \ = \ \sqrt{P(k)\left(d_{s,d}^{-\gamma}\right)}H_{s,d}(k)X(k) \ + \ N_{s,d}(k), \tag{2}$$

where $Y_{s,d}(k)$ indicates the signal received at the receiver on the '*k*th' subcarrier, $P(k)$ corresponds the power transmitted from the transmitter on the '*k*th' subcarrier, $d_{s,d}$ refers the distance between the transmitter and receiver, $H_{s,d}(k)$ denotes the channel coefficient between the transmitter and receiver on the '*k*th' subcarrier, γ represents the path loss exponent, $N_{s,\ d}(k)$ denotes the noise power contaminated on the transmitted signal with variance N_0.

So the total signal received at the receiver in direct transmission is represented as

$$Y_{s,d} = \textstyle\sum_{n=0}^{N-1} Y_{s,d}(k) \tag{3}$$

Similarly, during indirect transmission, the transmitted signal from the transmitter is received at the receiver through relay nodes. So the signal received at the relay on the '*k*th' subcarrier from the transmitter is expressed as

$$Y_{s,r}(k) \ = \ \sqrt{P(k) * (d_{s,\ r})^{-\gamma}}H_{s,\ r}(k)X(k) \ + N_{s,r}(k) \tag{4}$$

So the total signal received at the relay from the transmitter is represented as

$$Y_{s,r} = \textstyle\sum_{k=0}^{N-1} Y_{s,r}(k) \tag{5}$$

Also, the signal received at the receiver from the relay on '*k*th' subcarrier is expressed as

$$Y_{r,d}(k) \ = \ \sqrt{P(k) \ * \ (d_{r,\ d})^{-\gamma}}H_{r,d}(k)X(k) \ + \ N_{r,d}(k) \tag{6}$$

So the total signal received at the receiver from the relays on all the subcarriers is represented

$$Y_{r,d} \ = \ \textstyle\sum_{k=0}^{N-1} Y_{r,d}(k) \tag{7}$$

The 'SNR' between the transmitter and the receiver is calculated as

$$SNR_{s,d} \ = \ \textstyle\sum_{k=0}^{N-1} P(k)|H_{s,d}(k)|^2 \ /N_{s,d}(k) \tag{8}$$

Similarly the 'SNR' between the transmitter and relay is calculated as

$$SNR_{s,r} \ = \ \textstyle\sum_{k=0}^{N-1} P(k)|H_{s,r}(k)|^2 \ /N_{s,r}(k) \tag{9}$$

The ‘SNR’ between the relay and receiver is calculated as

$$SNR_{r,d} = \sum_{k=0}^{N-1} P(k)|H_{r,d}(k)|^{2} /N_{r,d}(k) \tag{10}$$

For amplify and Forward (AF) relay the amplification factor is expressed as

$$\beta = \sum_{n=0}^{N-1} P(k) /\sqrt{\sum_{n=0}^{N-1} P(k)|H_{s,r}(k)|^{2} + N_{0}} \tag{11}$$

The signal received at the receiver from the transmitter through the AF relay is expressed as

$$Y_{AF} = \beta * Y_{s,r} + Y_{r,d} \tag{12}$$

The total SNR between transmitter and receiver through ‘AF’ relay is calculated as

$$1/SNR_{AF} = 1/SNR_{s,r} + 1/SNR_{r,d} \tag{13}$$

So the channel capacity between the transmitter and receiver through AF relay is calculated as

$$Capacity_{AF} = \sum_{n=0}^{N-1} \sum (B/2)(1/N)\log(1 + SNR_{AF}) \tag{14}$$

Here ‘*B*’ represents the channel bandwidth of the considered network.

In case of DF relay, the total SNR between the transmitter and receiver through relay is calculated as

$$SNR_{DF} = argmin\,(SNR_{s,r}, SNR_{r,d}) \tag{15}$$

So the channel capacity between transmitter and receiver through DF relay is calculated as

$$Capacity_{DF} = \sum_{n=0}^{N-1} (B/2)(1/N)log(1 + SNR_{DF}) \tag{16}$$

At the receiver, for selection combining (SC) technique, the combined signal is expressed as

$$Y = argmax\,(SNR_{s,d}, SNR_{r,d}) \tag{17}$$

For MRC combining technique, the combined signal is expressed as

$$Y = \left(\sqrt{\sum_{n=0}^{N-1} P(k)} /N_{0}\right) Y_{s,d}H_{s,d}^{*} + \left(\sqrt{\sum_{n=0}^{N-1} P(k)} /N_{0}\right) Y_{r,d}H_{r,d}^{*} \tag{18}$$

4 Application of Relay Selection Algorithms for Performance Enhancement of MMR WiMAX Network

Multi-relay structure in a communication network is used to exploit the advantage of improvement in diversity gain as a result of which the received SNR is improved. To reduce the complexity and number of time slots taking for signal combining process, various relay selection schemes are proposed. By adopting these schemes, the more efficient relay path is selected which can improve the QOS parameters for the cooperative WiMAX network [6]. Here, two types of relay selection algorithms are applied for selecting the best relay path. These algorithms are (a) max_min based relay selection [11] (b) Harmonic mean of SNR-based relay selection [12]. The details of these algorithms are presented in the following subsections.

4.1 Max_Min Based Relay Selection Algorithm

```
Step1. Initialise: No. of relay 'R' = 4
       No. of subcarriers 'N' = 256
Step2. For j = 1: 4
       Compute $S_j = argmaxmin\left(SNR_{S,R_j}, SNR_{R_j,D}\right)$ using Eqs. (9) and (10).
Step3. Find out $P_{optimun}$ which represents the path having maximum SNR value
       from the vector '$S_j$'.
Step4. Estimate the values of $SNR_{S,R_j}$, $SNR_{R_j,D}$ of the optimum path from step 3.
Step5. Calculate channel capacity through AF and DF relay for the network using
       Eqs. (14) and (16) respectively.
       End.
```

4.2 Harmonic Mean of SNR-Based Relay Selection Algorithm

```
Step1. Initialise: No. of relay 'R' = 4
       No. of subcarriers 'N' = 256
Step2. For j = 1: 4
       Compute $S_j = \left(SNR_{S,R_j} * SNR_{R_j,D}\right) / \left(SNR_{S,R_j} + SNR_{R_j,D}\right)$ using Eqs. (9)
       and (10)
       End
```

Step3. Find out $P_{optimun}$ which represents the path having maximum harmonic mean of SNR value from the vector 'S_j'.

Step4. Estimate the values of SNR_{S,R_j}, $SNR_{R_j,D}$ of the optimum path from step 3.

Step5. Calculate the channel capacity through AF and DF relay for the network using Eqs. (14) and (16) respectively.

5 Simulation Results and Discussion

In this section, the impact of two relay selection schemes on the performance metrics, i.e., SER, channel capacity of the considered MMR WiMAX network are discussed. The simulation parameters used for the simulation are shown in Table 1. In Fig. 2a, b, the SER versus SNR performance comparison for various relay-diversity combining techniques with max-min and harmonic mean of SNR relay selection scheme has been demonstrated. From those figures, it is observed that for both the algorithms, DF-MRC scheme provides better SER performance as compared to other relay-diversity combining techniques. Again, considering DF-MRC case, at an SER level of 10^{-3}, the SNR requirement with max-min relay selection scheme is 12.2 dB. For the same condition, the SNR requirement with harmonic mean of SNR relay selection scheme is 8 dB. So with harmonic mean of SNR scheme at an SER level of 10^{-3}, an SNR improvement of 4.2 dB is observed as compared to max-min relay selection scheme. From Fig. 2c, it is observed that at 20 dB SNR, the channel capacities with harmonic mean of SNR relay selection algorithm in case of AF and DF relays are 8 Mbps and 11 Mbps, respectively. Similarly, for the same value of SNR in Fig. 2d, the channel capacities due to max-min scheme considering both AF and DF relays are found as 10 Mbps and 12 Mbps, respectively. So, at 20 dB SNR level, max-min relay selection scheme is superior to harmonic sum of SNR relay selection scheme by 2 Mbps and 1 Mbps with AF and DF relay respectively.

Table 1 Simulation parameters

Parameter	Value
WiMAX transmitter power	43dBm
Channel bandwidth	5 MHz
Transmission technique	OFDMA
Total no.of subcarriers	256
No.of data subcarriers	192
No.of pilot subcarriers	8
No.of guard band subcarriers	56
No.of subcarriers for cyclic prefix	64
Channel model	Rayleigh
Path loss exponent	3

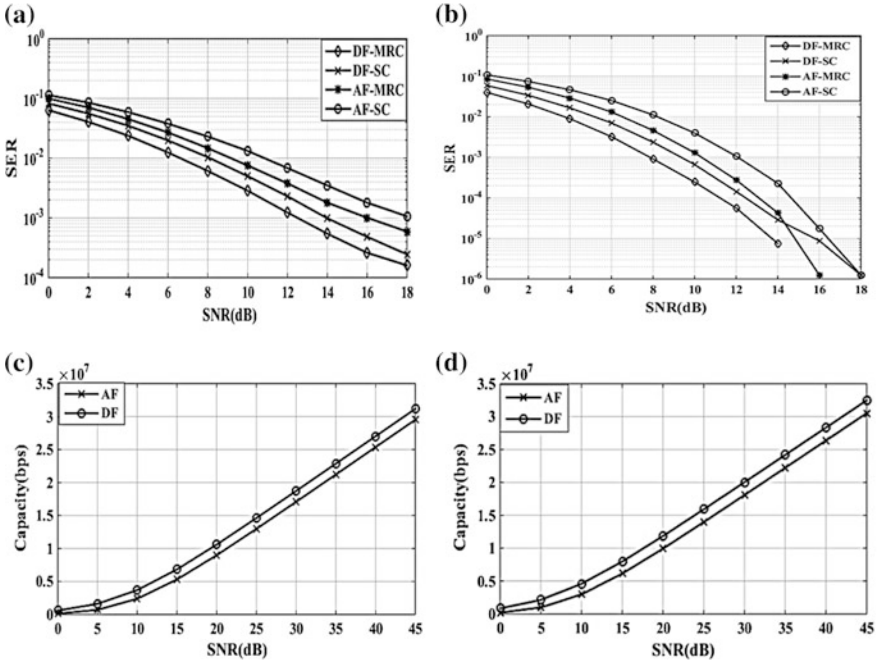


Fig. 2 **a** and **b** SER versus SNR characteristics curve of max-min and harmonic mean of SNR based relay selection schemes respectively. Figure 2**c** and **d** Channel capacity due to harmonic mean of SNR and max-min based relay selection schemes respectively

6 Conclusion

In this paper, simulation-based investigations are carried out and the effect of different relay selection algorithms on the performance of various QOS parameters like SER, channel capacity on the considered MMR WiMAX network are studied. It is observed that the SER performance of DF-MRC combination outperforms all other relay protocol-diversity combining techniques using both the relay selection schemes. Again comparing the two algorithms, the harmonic mean of SNR relay selection algorithm outperforms max-min relay selection algorithm in providing less SER. Again the max_min relay selection algorithm provides better channel capacity as compared to harmonic mean of SNR algorithm for both AF and DF relay case. The work carried out can be extended for point to multipoint (PMP) MMR WiMAX network to resemble real-network scenario.

References

1. D. Soldani and S. Dixit., “Wireless relays for broadband access.,” *IEEE communications Magazine,* vol. 46, no. 3, pp. 58–66, 2008.
2. Y. Yang, H. Hu, J. Xu and G. Mao, “Relay Technologies for WiMAX and LTE-Advanced Mobile Systems,” *IEEE Communications Magazine,* vol. 47, no. 10, pp. 100–105, 2009.
3. C. Nie, P. Liu, T. Korakis, E. Erkip and S. S. Panwar, “Cooperative Relaying in Next Generation Mobile WiMAX Networks,” *IEEE Transactions on Vehicular Technology.,* vol. 62, no. 3, pp. 1399–1405, 2013.
4. Y. Ge, S. Wen, Y. H. Ang and Y. C. Liang, “Optimal Relay Selection in IEEE 802.16j Multihop relay Vehicular Networks,” *IEEE Transactions on Vehicular Technology,* vol. 59, no. 5, pp. 2198–2206, 2015.
5. S. Ann and H. S. Kim, “Relay association method for optimal path in IEEE 802.16j mobile,” *European Transactions on Telecommunications,* vol. 21, no. 7, p. 624–631, 2010.
6. A. S. Ibrahim, A. K. Sadek, W. Su and K. Liu, “Relay selection in Multinode Cooperative Communication: When to cooperate and Whom to cooperate with?,” in *IEEE Globecom Proceeding*, Buffalo, USA, 2006.
7. T. T. Duy and H. Y. Kong, “Decode and Forward Relaying Systems with Nth Best Relay Selection over Rayleigh Fading Channels,” *Journal of the Korean Institute of Electromagnetic engineering and Science,* vol. 12, no. 1, pp. 8–12, 2012.
8. L. R. Lakshmi, “Link Quality Based Path Selection Methods for IEEE 802.16j Mobile Multihop Relay Networks,” *Wireless Personal Communications,* vol. 75, no. 5, pp. 579–592, 2014.
9. H. Lee, H. Hwang, S. Kim, B. Roh and G. Park., “Cooperative RS Selection Schemes for IEEE 802.16j Networks,” in *IEEE Military Communications Conference.*, Seoul, Korea, 2013.
10. P. Mangyarkarasi and S. Jayashri, “Efficient relay selection scheme using regenerative and degenerative protocols for 5th generation WiMAX systems,” *International Journal of Electronics and Communication* (*AEU*), vol. 69, no. 5, pp. 781–789, 2015.
11. I. Krikidis, J. S. Thompson, S. McLaughlin and N. Goertz, “Max-Min Relay Selection for Legacy Amplify-and-Forward Systems with Interference.,” *IEEE Transactions on Wireless Communications,* vol. 8, no. 6, pp. 3016–3027, 2009.
12. Y. Jing and H. Jafarkhani, “Single and Multiple Relay Selection Schemes and their Achievable Diversity Orders,” *IEEE Transactions on Wireless Communications,* vol. 8, no. 3, pp. 1414–1423, 2009.

Predicting Link Failure in Vehicular Communication System Using Link Existence Diagram (LED)

Sourav Kumar Bhoi, Munesh Singh and Pabitra Mohan Khilar

Abstract Link failure is a major problem in vehicular communication system because of high mobility of vehicles in the network. By predicting the link failure at an early stage the performance of the system can be improved. In this paper, we proposed a data forwarding technique by predicting the link failure in a route by generating a Link Existence Diagram (LED). LED shows whether a link exist between the vehicles or not. Second, we have calculated the Link Expiration Time (LET) between the two vehicles in a route. Third, we have generated a *Time-Link* diagram to synchronize LET with the data transmission time. Finally, by updating the LET and checking the existence of LET, we generate the LED. Simulations are performed to evaluate the performance of the proposed protocol.

Keywords VANET · Routing · LET · LED · Time-Link diagram · Vehicle-Position diagram

1 Introduction

Link failure degrades the performance of the system by hindering the data communication process. In VANET, the main reason of link failure is speed of the vehicles in the city areas, which disrupts the wireless connection between the vehicles [1–3]. The main objective of implementing VANET is to provide safety and comfort services to the drivers and passengers. To send the data through a route, an optimal

S.K. Bhoi (✉) · M. Singh · P.M. Khilar
Department of Computer Science & Engineering, National Institute of Technology,
Rourkela 769008, Odisha, India
e-mail: souravbhoi@gmail.com

M. Singh
e-mail: muneshpal24@gmail.com

P.M. Khilar
e-mail: pmkhilar@nitrkl.ac.in

P.K. Sa et al. (eds.), *Progress in Intelligent Computing Techniques: Theory, Practice, and Applications*, Advances in Intelligent Systems
and Computing 719, DOI 10.1007/978-981-10-3376-6_54

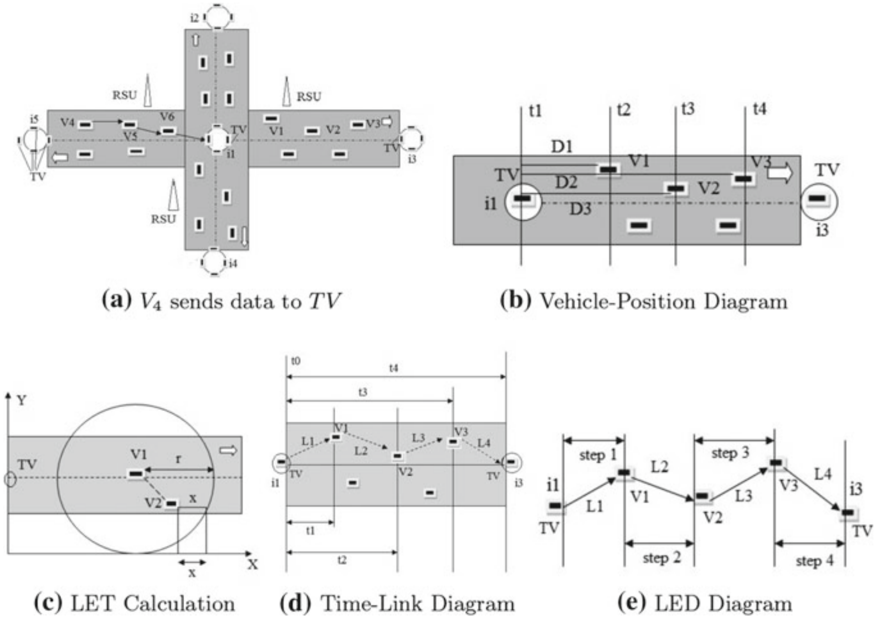
routing protocol is required, however it suffers from link breakage problem. If link breakage is encountered at the data forwarding phase then many protocols use carry and forward mechanism. But in the proposed model, we predict the link failure at an early stage and forward the data in that route which has less link breakage problems. We have generated LED diagram to predict the link failures. Many works have been proposed in prediction-based routing to deliver the data successfully to the destination. Schoch et al. [4] surveyed on the different communication patterns which helps in VANET routing. Menouar et al. [5] proposed a Movement-Based Routing Protocol (MOPR) to predict the positions of the vehicles by considering static speed and time. Namboodiri et al. [6] proposed a Prediction-Based Routing Protocol (PBR) to predict the route lifetimes and select another route before the lifetime of a route expires. Kong et al. [7] proposed a solution to estimate the traffic in urban roads using GPS data. Xue et al. [8] proposed a Prediction-Based Soft Routing Protocol (PSR) based on Vehicular Mobility Pattern (VMP) trace taken from Sanghai city.

The remaining portion of the paper is described as follows. Section 2 presents the system model and functionality of the proposed protocol. Section 3 focuses on the simulation and results. Section 4 presents a brief conclusion.

2 System Model and Functionality

In this system, we have considered many assumptions. The city is considered as a graph with intersections (i) as vertices and roads connecting the intersections as edges. Vehicles (V) are installed with GPS service and maps. Vehicle sends position information to the neighbor vehicles at a particular interval. For routing, vehicles use MFR (Most Forward within Radius) routing protocol [9]. Destination (D) position is known to the source vehicle and it is updated by the Trusted Vehicles (TV) which are fixed at the intersections (using location services). There is a less chance of link breakage between the vehicles in the city area, but sometimes vehicles are far apart to establish link. The positions of the vehicles generated in the *Vehicle-Position* diagram are considered as the actual positions of the vehicles with a negligible error rate. We have considered only transmission time (τ_t) with negligible other times (τ_o) like processing time, queuing time, and propagation time. Therefore, the time to send the data from one vehicle to other is denoted by $(\tau_t + \tau_o)_{V_1V_2}$.

In this system, every intersection is set with fixed Trusted Vehicles (TV) to forward the data in a selected direction. If there are p directions then there are p number of TV. Vehicles with the data packets when reaches at the intersection, handovers the data to TV. TV selects another vehicle in the calculated shortest path (SP) to send the data to D. V2V (Vehicle-to-Vehicle) communication plays an important role in forwarding the data to D. The vehicle which receives the data packet from TV is unaware about the further link conditions. If TV is aware about the link conditions ahead at an early stage then it forwards the data in that route which is resilient against the link failure problems. TV stores many information like vehicle ID, Speed (S), Average Speed (AS), Location (L), Vehicle Passing Time through the intersection



(a) V_4 sends data to TV **(b)** Vehicle-Position Diagram

(c) LET Calculation **(d)** Time-Link Diagram **(e)** LED Diagram

Fig. 1 Generation of LED

(VPT), Direction (Dir), SP, Incoming Vehicles/Outgoing Vehicles (IV/OV), LED and Link Existence Status (LES). ID, S, AS, and L are provided by beaconing service. The rest information is captured and calculated by TV using sensor data and high computing systems. These data are updated by TV at a particular interval of time for every neighboring intersection. Only LEDs are generated on demand for a neighboring intersection. In Fig. 1a, we see that TVs are set at every intersection. The nearest intersections to i_1 are i_2, i_3 and i_4. From Fig. 1a, after V_4 initializes the communication it transmits the data to TV_{i_1} using MFR routing protocol in the path V_4-V_5-V_6-TV_{i_1}. Now TV_{i_1} calculates SP to D because D is moving and let the path be TV_{i_1}-V_1-V_2-V_3-TV_{i_3}-...-TV_{i_n}-D where $n = 1, 2, ..., n$. After this TV_{i_1} checks the LED between i_1 and i_3 to confirm whether data transmission is possible.

To generate LED, TV_{i_1} generates an approximate $Vehicle - Position$ diagram. As we know TV_{i_1} awares about the average speed AS of a vehicle and it knows VPT (time at which a vehicle crosses the intersection), it can calculate the approximate positions of the vehicles at a particular instant of time. From Fig. 1b the distance D_1, D_2 and D_3 are calculated by ($AS * \Delta t$) where Δt is the time difference (t_2-t_1). By this TV_{i_1} gets the approximate vehicles position at a particular instant of time. After that TV_{i_1} finds a route for reference (R_f) and checks the link problem in this route. Using the $Vehicle - Position$ diagram, TV_{i_1} can find a route using MFR routing protocol according to the range of the vehicles. Let the route calculated is found to be $R_f = TV_{i_1}$-V_1-V_2-V_3-TV_{i_3}. Now TV_{i_1} checks the link existence of the whole route R_f by calculating the LET. LET is calculated between the two vehicles to find the rest connection time between them. From *Vehicle – Position* diagram if we consider V_1 and V_2 and assume that V_2 is faster than V_1 then from Fig. 1c V_2 is out of the range after x distance and link breaks. LET is calculated to be:

$$LET = \frac{x}{AS_{V_2} - AS_{V_1}} \tag{1}$$

The LET calculated in Eq. 1 is updated by updating *Vehicle – Position* diagram. First, TV_{i_1} calculates LET between itself and V_1 and it is called as LET_1. Then it calculates LET_2 between V_1-V_2, LET_3 between V_2-V_3 and LET_4 between V_3-TV_{i_3}. After this TV_{i_1} synchronizes the time of data transmission with LET timings between the vehicles. This synchronization states that if we transmit the data to TV_{i_3} then what is the chance that after sending it to V_1 from TV_{i_1} link L_2 (link between V_1 and V_2) exists. This is analyzed from a *Time – Link* diagram. The *Time – Link* diagram is created and shown in Fig. 1d. Let the transmission time from one vehicle to other vehicle is considered to be $\left(\tau_t + \tau o\right)_{V_1 V_2}$ where τ_t is the approximate transmission time (calculated by considering $\frac{Packet\ Size}{Bandwidth}$) from one vehicle to other vehicle and τ_o is the negligible other delays (*processing delay* + *propagation delay* + *queuing delay*). Steps for the generation of LED diagram is described as follows:

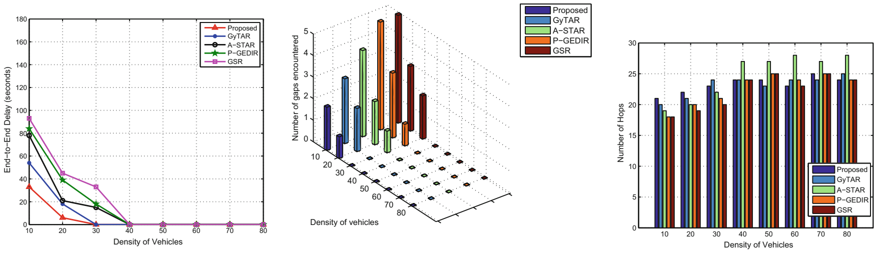
- Step 1: If data is send at t_0 time then check if $LET_1 < \left(\tau_t + \tau o\right)_{TV_{i_1} - V_1}$ and if this condition is true then L_1 exists. Figure 1e shows the first step of LED diagram.
- Step 2: If data is send after $t_1 = \left(\tau_t + \tau o\right)_{TV_{i_1} - V_1}$ time, then check whether updated LET_2 exists ($LET_2 > 0$) and if $LET_2 > \left(\tau_t + \tau o\right)_{V_1 - V_2}$ then L_2 exists. Figure 1e shows the second step of LED diagram.
- Step 3: If data is send after $t_2 = \left(\tau_t + \tau o\right)_{TV_{i_1} - V_1} + \left(\tau_t + \tau o\right)_{V_1 - V_2}$ time, then check whether updated LET_3 exists ($LET_3 > 0$) and if $LET_3 > \left(\tau_t + \tau o\right)_{V_2 - V_3}$ then L_3 exists. Figure 1e shows the third step of LED diagram.
- Step 4: If data is send after $t_3 = \left(\tau_t + \tau o\right)_{TV_{i_1} - V_1} + \left(\tau_t + \tau o\right)_{V_1 - V_2} + \left(\tau_t + \tau o\right)_{V_2 - V_3}$ time, then check whether updated LET_4 exists ($LET_4 > 0$) and if $LET_4 > \left(\tau_t + \tau o\right)_{V_3 - TV_{i_3}}$ then L_4 exists. Figure 1e shows the fourth step of LED diagram.

The final LED diagram is generated after step four and from this TV_{i_1} decides whether link exists or not and assign $LES = 1$ (if link exists $LES = 1$ else $LES = 0$). After getting the LES information TV_{i_1} decides whether to calculate a new route to forward the data or send the data in the same route. For a dynamic scenario (real-life scenario) the link breakage occurs between the vehicles in each direction. By analyzing the LED diagrams in each direction, if all directions have link breakage problem then we send the data in that direction which has a less approximated transmission time.

3 Simulations and Results

MATLAB R2015a is used to simulate the performance of the scheme. The proposed routing protocol is compared with the standard existing routing protocols. These protocols work well in the city scenarios. End-to-End delay is the main parameter to evaluate the performance. We have also considered parameters like communications gaps and path length. The simulation environment is set to $3000 \times 3000\,\mathrm{m}^2$. The road distance is 1000 m. The number of vehicles in the roads are generated using Poisson distribution. The 2-D positions of the vehicles in both directions in a road are generated using uniform random distribution. The positions are assumed to be the current positions. The communication range is set to 200 m. The data rate is set to 3 Mbps. 512 bytes packet size is used and the individual vehicle speed is randomly generated between 30 to 50. The simulation is run 100 times to find the average of the performance metrics. Source vehicle and destination are selected randomly. The data is assumed to be received at the receiver end, if it is in the range. The Manhattan mobility model is used for simulation.

Figure 2a shows that proposed method shows better results than other routing protocols. When the density is low say 10 and 20, proposed method sends the data in a minimum time than the other routing protocols because it sends the data through that path which is highly connected. Figure 2b shows that proposed method shows better results than other routing protocols. When the density is low say 10 and 20, proposed method shows less gaps than the other routing protocols because it sends the data through that path which is highly connected. Figure 2c shows that proposed method shows better results than other routing protocols. When the density is low say 10 and 20, proposed method uses shows more hops than the other routing protocols because it sends the data through that path which is highly connected.



(a) End-to-End delay **(b)** Number of gaps **(c)** Number of hops

Fig. 2 Performance analysis of the proposed protocol

4 Conclusion

In this paper, the links between the junctions are predicted using the LED. This increases the performance of the system by reducing the end-to-end delay. This strategy will be a better routing solution for VANET applications. Simulations show that proposed method overcomes the drawbacks of existing schemes.

Acknowledgements We want to thanks Computer Science and Engineering Department, NIT, Rourkela, India for providing infrastructure to conduct the experiments.

References

1. F., Li. and Wang, Y.: Routing in vehicular ad hoc networks: A survey. IEEE Vehicular Technology Magazine, 2(2), 12–22 (2007)
2. Zeadally, S., Hunt, R., Chen, Y.-S., Irwin, A., and Hassan, A.: Vehicular ad hoc networks (VANETS): status, results, and challenge. Telecommunication System, 50(4), 217–241 (2010)
3. Nzouonta, J., Rajgure, N., Wang, G., and Borcea, C.: VANET routing on city roads using real-time vehicular traffic information. Vehicular Technology, IEEE Transactions on. 58(7), 3609–3626 (2009)
4. Schoch, E., Kargl, F., Weber, M., and Leinmuller, T.: Communication patterns in VANETs. Communications Magazine, IEEE, 46(11), 119–125 (2008)
5. Menouar, H., Lenardi, M., and Filali, F.: Movement Prediction-Based Routing (MOPR) Concept for Position-Based Routing in Vehicular Networks. Vehicular Technology Conference, IEEE 66th, 2101–2105 (Oct. 2007)
6. Namboodiri, V., and Gao, L.: Prediction-based routing for vehicular ad hoc networks. IEEE Transactions on Vehicular Technology, 56(4), 2332–2345 (2007)
7. Kong, Q. J., Zhao, Q., Wei, C., and Liu, Y.: Efficient Traffic State Estimation for Large-Scale Urban Road Networks. IEEE Transactions on Intelligent Transportation Systems, 14(1), 398–407 (2013)
8. Xue, G., Luo, Y., Yu, J., and Li, M.: A novel vehicular location prediction based on mobility patterns for routing in urban VANET. EURASIP Journal on Wireless Communications and Networking, 1–14 (2012)
9. Raw, R., and Lobiyal, D.: B-MFR routing protocol for vehicular ad hoc network. In Networking and Information Technology (ICNIT), International Conference on, 420–423 (2010)

Part VIII
Communication Systems, Antenna Research, and Cognitive Radio

An M-Shaped Microstrip Antenna Array for WLAN, WiMAX and Radar Applications

Aastha Gupta, Vipin Choudhary and Malay Ranjan Tripathy

Abstract A compact multiband antenna, using HFSS, is designed on FR4_epoxy substrate. The dimension of substrate are chosen to be 50 × 75 × 1.6 mm^3 and has a permittivity of 4.4. An M-shaped antenna with symmetrical slots is used to form an array which is proposed in this paper. Inset feeding is made to make it simple and compact. Multibands are obtained in S11 versus frequency plots. The effects of forming an array of the antenna elements had been evaluated in order to obtain an upgraded performance of the presented antenna in the WLAN, WiMAX and Radar range. Lower bands were obtained with operating frequencies at 2.4, 3.6, and 5 GHz suitable for WLAN and WiMAX applications. For the operations in K-band Radar range, interesting results were obtained in 11.3–14.4 GHz band.

Keywords Antenna array · WLAN · WiMAX · Radar

1 Introduction

One of the fastest growing domain in present arena is the wireless communication engineering and its use for simultaneous multiple activities. To support such dynamic and high speed communications, multiband with high gain and wide band systems are required. A lot of efforts are being made to design efficient and high-gain planar antennas to facilitate such applications. Microstrip patch antennas with its valuable advantages of being low profile, cheap, and of low weight have given a boost to the wireless communication field such as Wireless Local Area

A. Gupta (✉) · V. Choudhary · M.R. Tripathy
Department of Electronics and Communication Engineering,
Amity University, Noida 201313, UP, India
e-mail: gupta.aastha16@gmail.com

V. Choudhary
e-mail: choudhary.vip77@gmail.com

M.R. Tripathy
e-mail: mrtripathy@amity.edu

P.K. Sa et al. (eds.), *Progress in Intelligent Computing Techniques: Theory, Practice, and Applications*, Advances in Intelligent Systems and Computing 719, DOI 10.1007/978-981-10-3376-6_55

Networks, Satellite, and radar communication, etc. With increase in long distance wireless communication requirement, an important aspect concerning the fabrication of the antennas with high gain and high directivity is the need of the hour. Antenna arrays of mircostrip patch antenna enhances the signal strength in the desired direction, thus improving the antenna parameters.

The trends and applications of adaptive antenna are discussed in a paper [1]. For beam forming applications Yoshida et al. [2] proposed array antenna based on 3-D SiP structure in small wireless terminals. The conformal array antenna based on millimeter-wave-shaped beam substrate integrated structure is reported in a paper [3]. For different applications in WLAN, WiMAX, multiband, W band various array antennas with unique designs are proposed in the literature [4–8, 11]. Near-field imaging application is proposed using array antenna in the paper [9]. Aguilar et.al. used array antenna for space application [10].

This paper provides an analysis of different array configurations of a single M-shaped patch antenna. It was observed that on increasing the array elements, better gain characteristics were obtained. An antenna array configuration was found suitable for the WLAN, WiMAX, and K-band radar applications. Multibands resulting with different values of return loss, along with different gains and impedance bandwidth are obtained from different designs.

The paper is organized as follows. Section 2 illustrates the antenna design and its characteristics. Results and discussion are depicted in the Sect. 3. Conclusion is made in Sect. 4.

2 Paper Preparation

Based on the frequency at which we require our antenna to operate, the parametric values were chosen. Figure 1a shows an M-shaped antenna with two symmetric rectangular slots on the patch and triangular slots on the feed (Design A). The substrate used has a thickness of h = 1.6 mm, permittivity Ɛr = 4.4 and dimensions of 50 × 75 mm^2. For a prime performance, circuit impedance changes must be minimum, which require uniform isotropic dielectric constant. The dielectric constant of the substrate should lie between 2.2–12 so as to have a larger bandwidth.

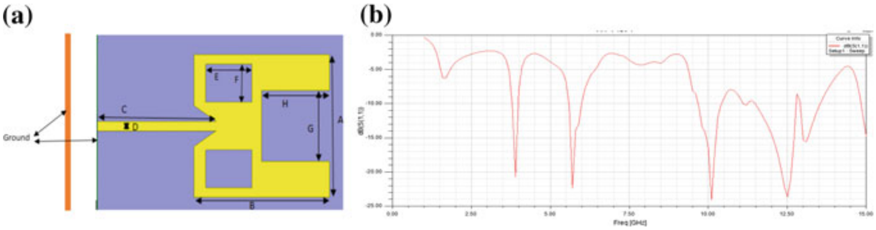


Fig. 1 **a** Design of single element monopole patch antenna, and **b** Return loss versus frequency plot

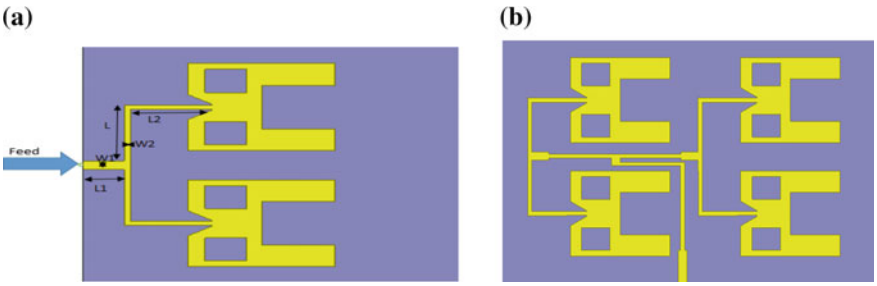


Fig. 2 **a** 2 × 1 array antenna design (Design B) and, **b** 2 × 2 array antenna design (Design C)

The parameters set for the patch are as follows: A = 30 mm, B = 35 mm, C = 30.8 mm, D = 2 mm, E = 12 mm, F = 8 mm, G = 15 mm and H = 17.5 mm. The feed provided to this design is an inset feed with a feed line width (C) of 30.8 mm and a feed line inset distance (D) of 2 mm. This antenna operated at frequency of 10 GHz, 12.5 GHz and 13 GHz with an application area of X-band and Ku-band.

Further, in order to improve gain characteristics, a 2 × 1 array (Design B) of M-shaped patch antenna was formed. Design B is proposed with a change in dimensions of the two symmetric rectangular slots to 8 × 10 mm^2. The feed given is an inset feed. The resultant design is shown in Fig. 2 with feed dimensions set as follows: L = 19.2 mm, L1 = 11.41 mm, L2 = 20.7 mm, W1 = 3.05 mm, and W2 = 1.43 mm. The distance between the two elements in the 2 × 1 array is 10 mm. This multiband antenna design shows interesting results at 3 GHz, 5.8 GHz and 11.3 GHz, suitable for Wlan, WiMAX, and Radar applications.

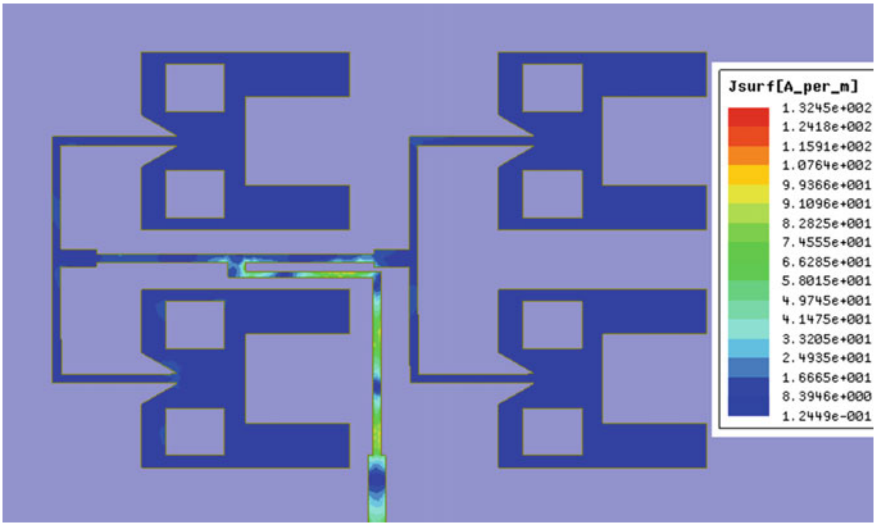


Fig. 3 Current distribution for design C at 5 GHz

Table 1 Dimensions of 2 × 1 and 2 × 2 array antenna

Parameters	Design B	Design C
Length of the patch (Single element)	35 mm	35 mm
Width of the patch (Single element)	30 mm	30 mm
Substrate material	FR4_epoxy	FR4_epoxy
Dielectric constant	4.4	4.4
Substrate thickness	1.6 mm	1.6 mm
Length of the ground	90 mm	150 mm
Width of the ground	90 mm	90 mm
Feeding Technique	Inset	Inset

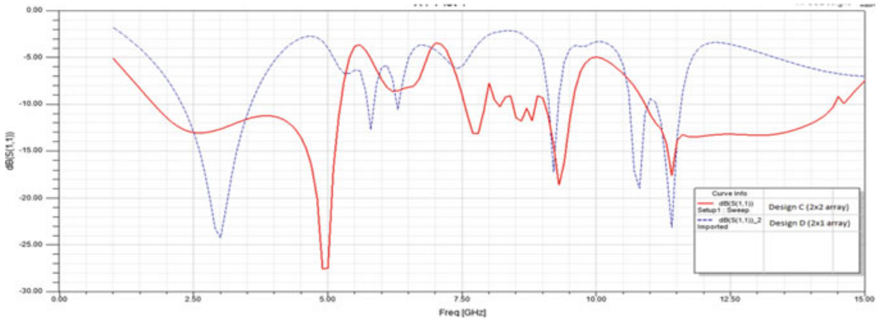


Fig. 4 The Return loss versus frequency graph pf design B and design C

Table 2 Return loss, bandwidth and gain results for design A, B and C

Design	S_{11} dB (at *f* GHz)	Δf (Bandwidth) (MHz)	Gain (Max)
A	–24 dB(10 GHz)	900	2
	–23.6 dB(12.5)	1300	7.9
	–15.5 dB(13 GHz)	600	16.3
B	–24.3 dB(3 GHz)	1400	4
	–12.6 dB(5.8 GHz)	100	5.5
	–23 dB(11.3 GHz)	400	13
C	–27.5 dB(4.8 GHz)	3500	23.5
	–18.6 dB(9.3 GHz)	500	4.1
	–17.5 dB(11.3 GHz)	3500	20.6

Another step was taken, with an antenna Design C that is outlined by forming a 2 × 2 array of Design A, with a change in symmetric rectangular slots to 8 × 10 mm^2. Figure 3 depicts Design C, which also is having an inset feed. The simulation results shows that this design of 2 × 2 array radiates magnificently at 2.4, 3.6, 5, and 12.5 GHz having wide application in WLAN, WiMAX, and K-band Radar. Table 1 shows the various design parametric values for design B and design C.

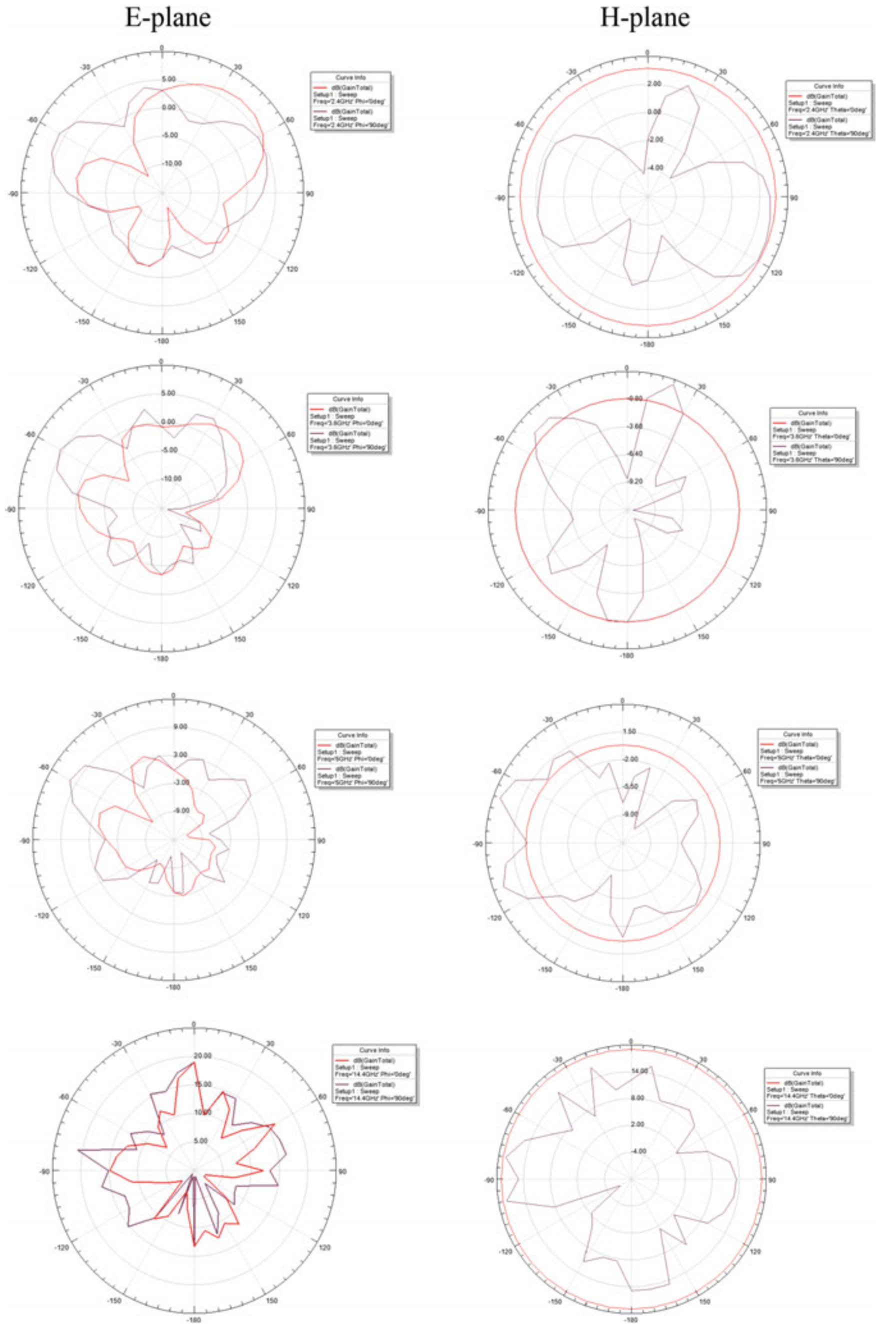


Fig. 5 E-plane and H-plane radiation pattern for design C at 2.4, 3.6 and 5 GHz

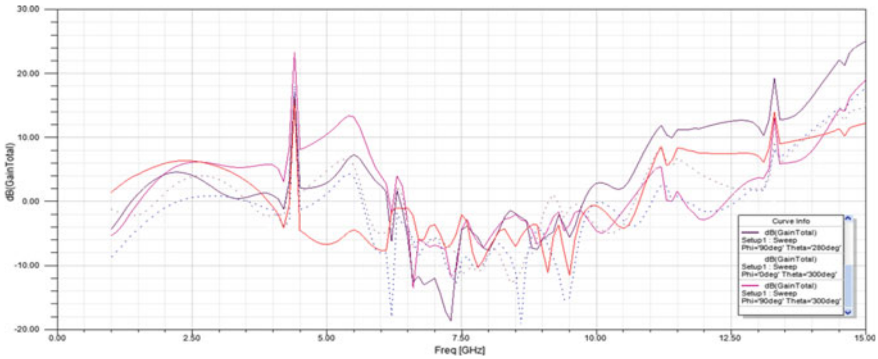


Fig. 6 Gain versus Frequency plot of design C

In order to understand the performance of the proposed 2 × 1 and 2 × 2 array antennas, with the help of HFSS, the current distribution for these designs was analyzed. The current distribution helps to print the antennas on the substrate to produce low-cost and repeatable antennas in low profile. The current distribution for design C is depicted in Fig. 3.

3 Results and Discussion

In this section, various simulation results are discussed and studied upon. The return loss versus frequency plot for the proposed 2 × 1 and 2 × 2 array antennas is shown in Fig. 4. Design B and design C is found to be useful for WLAN, WiMAX, and Radar operations. The return loss values for design B are –24.3 dB, –12. 6 dB, and –23.1 dB obtained at 3 GHz, 5.8 GHz, and 11.3 GHz with respective bandwidths of 1400 MHz, 100 MHz and 400 MHz and band gain of 4 dB, 5.5 dB, and 13 dB. Also for design C, in the return loss versus frequency plot, the useful peaks are obtained at 4.8 GHz, 9.3 GHz, and 11.3 GHz with return loss of –27.5 dB, –18.6 dB, and –17.5 dB in the respective bandwidth of 3500 MHz, 500 MHz, and 3500 MHz. Table 2 depicts various result parametric values of design A, B, and C.

Shown in the Fig. 5 are the simulated results for E-plane and H-plane radiation patterns for Design C (2 × 2 array) at frequencies 2.4, 3.6, 5, and 14.4 GHz. From the omnidirectional E-plane radiation pattern at 2.4, 3.6, 5, and 14.4 GHz, it can be inferred that Design C can be widely used for WLAN and WiMAX applications. Also the radiation pattern in H-plane is stable enough.

Gain versus frequency plot of Design C shows interesting results that are widely applicable in WLAN and WiMAX range. The gain values obtained are 6.5 dB, 5.5 dB, 10.5 dB, and 20.6 dB at 2.4 GHz, 3.6 GHz, 5 GHz, and 14.4 GHz with the return loss of –12.9 dB, –11.4 dB, –27.5 dB, and –10.5 dB, respectively. Figure 6 shows the gain versus frequency plot of design B and design C.

4 Conclusion

In this paper, a highly efficient design of antenna has been proposed. The antenna array showed magnificent results in comparison with the single patch antenna. The resultant characteristics of the proposed antenna made it suitable for WLAN applications such as in-home networks, hotspots in hotels, campuses, etc., as well as WiMAX and Radar applications. It worked on the resonant frequencies of 2.4, 3.6, 5, and 14.4 GHz.

References

1. Chandran S, "Adaptive antenna arrays: trends and applications." Springer Science & Business Media; (2013).
2. Yoshida S, Suzuki Y, Ta TT, Kameda S, Suematsu N, Takagi T, Tsubouchi K, "A 60-GHz band planar dipole array antenna using 3-D SiP structure in small wireless terminals for beamforming applications." Antennas and Propagation, IEEE Transactions; 61(7): (2013) 3502-10.
3. Cheng YJ, Xu H, Ma D, Wu J, Wang L, Fan Y, "Millimeter-wave shaped-beam substrate integrated conformal array antenna." Antennas and Propagation, IEEE Transactions; 61(9): (2013);4558-66.
4. Falahati A, Naghshvarian Jahromi M, Edwards RM. "Wideband fan-beam low-sidelobe array antenna using grounded reflector for DECT, 3G, and ultra-wideband wireless applications." Antennas and Propagation, IEEE Transactions; 61(2); (2013);700-6.
5. Rafii V, Nourinia J, Ghobadi CH, Pourahmadazar J, Virdee BS. "Broadband circularly polarized slot antenna array using sequentially rotated technique for-band applications." Antennas and Wireless Propagation Letters, IEEE; 12: (2013); 128-31.
6. Cheng YJ, Guo YX, Liu ZG. "W-Band large-scale high-gain planar integrated antenna array." Antennas and Propagation, IEEE Transactions. 62(6): (2014); 3370-3.
7. Yeo J, Lee JI. "Broadband series-fed two dipole array antenna with an integrated balun for mobile communication applications." Microwave and Optical Technology Letters. 1; 54(9): (2012):2166-8.
8. Chitra RJ, Nagarajan V. "Double L-slot microstrip patch antenna array for WiMAX and WLAN applications." Computers & Electrical Engineering. 30; 39(3): (2013) 1026-41.
9. Tofigh F, Nourinia J, Khazaei KM. "Near-field focused array microstrip planar antenna for medical applications." Antennas and Wireless Propagation Letters, IEEE. 13; (2014): 951-4.
10. Aguilar AG, Jakobus U, Kulkarni SR. "Antenna analysis for space applications with the electromagnetic field solver FEKO." In Space Science and Communication (IconSpace), 2015 International Conference. IEEE (2015) (pp. 398–401).
11. Wang L, Guo YX, Sheng WX. "Wideband high-gain 60-GHz LTCC L-probe patch antenna array with a soft surface." Antennas and Propagation, IEEE Transactions. 61(4); (2013):1802-9.

Gain Enhancement of Microstrip Patch Antenna Using H-Shaped Defected Ground Structure

Asmita Rajawat, P.K. Singhal, Sindhu Hak Gupta, Chavi Jain, Praneet Tomar and Kartik Kapur

Abstract This paper presents the enhancement of the gain of an inset feed microstrip slot antenna using H-shaped Defected Ground Structure (DGS). Reduced higher order harmonics increases the gain of the antenna using DGS technology. A critical comparative analysis is made between the antenna with and without DGS. Simulation results reveal that there is a remarkable increase in gain. The proposed design is applicable for commercial frequency band of 900 MHz–2.4 GHz.

Keywords Inset feed • Microstrip patch • H-shaped Defected Ground Structure

A. Rajawat (✉) · S.H. Gupta · C. Jain · P. Tomar · K. Kapur
Amity University, Noida, Uttar Pradesh, India
e-mail: arajawat@amity.edu

S.H. Gupta
e-mail: shak@amity.edu

C. Jain
e-mail: chavi229@gmail.com

P. Tomar
e-mail: praneet.tomar13@gmail.com

K. Kapur
e-mail: kartikmart17@gmail.com

P.K. Singhal
MITS, Gwalior, India
e-mail: pks_65@yahoo.com

P.K. Sa et al. (eds.), *Progress in Intelligent Computing Techniques: Theory, Practice, and Applications*, Advances in Intelligent Systems and Computing 719, DOI 10.1007/978-981-10-3376-6_56

1 Introduction

1.1 Related Work

With the increasing technology, antennas have become an important part in that they allow users to transmit and receive data to communicate with the infrastructure.

Microstrip patch antennas are widely used in today's world in microwave and wireless communication system because of light weight, low-fabrication cost, low profile, easy fabrication, and ability to conform to any shape. In this modernizing era, people rely on the wireless Internet on a daily basis and thus here, a suitable design for an antenna is chosen so as to work in the commercial range of 900 MHz–2.4 GHz frequencies. This antenna is further modified to enhance gain by introducing the Defected Ground Structure.

Different techniques of enhancement of gain are reported such as array configuration [1], using EBG [2], use of parasitic elements technology [3], etc. In this paper, DGS is introduced for gain enhancement of a microstrip slot antenna.

Defected Ground Structure (DGS) [4–6] refers to a configuration that can be an etched periodic or cascaded nonperiodic configuration. It creates a defect in the ground of transmission line. This transmission line maybe of any kind which produces disturbance in the distribution of shield current in the ground plane. Some characteristics changes of the transmission line can be seen because of the defect such as capacitance and inductance. So, it can be said that these characteristics such as effective capacitance and inductance can be increased due to the defect. DGS has created a degree of freedom in microwave designing and has opened the door to multiple applications. Also, it has become an interesting area of research because of their extensive applicability. Introducing DGS in our antenna design has brought significant change in the parameters of the design and the results have been analyzed. In our research, DGS helps in increasing the gain parameter of the designed antenna.

1.2 Contribution

The research contribution of the paper is dedicated toward the study and simulation of three different antenna designs. The design is modified by using a novel H-shaped Defected Ground Structure. The design concept of the paper is based on an inset feeding technique because of its ease of fabrication, simple matching technique (because of the adjustment in its inset position). It is shown that the first antenna designed without DGS has a lower gain and the gain increases by adding a defect in the ground which further increases by addition of one more defect in the ground.

1.3 Organization of Paper

The paper is organized as follows. Section 2 and Sect. 3 presents the proposed antenna designs and its simulation results, which validates the accuracy of the analysis and demonstrates the strategy for different antenna designs with and without DGS. Section 4 concludes the paper.

2 Antenna Design

The paper presents three new microstrip patch antenna designs. The software used for designing of the antenna is CST (Computer Simulation Technology) [7] and the chosen value of the frequency in the task for simulation is 2.45 GHz. First, an antenna is crafted on a 1.6 mm thick FR-4 (lossy) substrate having a dielectric constant of 4.3 which is generally taken high because it reduces the size of the antenna and makes it compact. This design consists of a main patch and two slots cut on the main patch. The second antenna employs Defected Ground Structure using H-Shape for enhancing the gain of the antenna. The third antenna also employs DGS with two more slots cut on the ground. The antenna designs are obtained for various parameters and based on the results obtained, a comparison is made among them.

2.1 Antenna Design A

Figure 1 shows a model of the antenna and named as design A with two slots and does not contain DGS. The newly crafted model design is provided with an inset feed [8, 9] in the centre of the patch and gain and return loss are measured at 2.45 GHz frequency.

Measurements of patch are determined using the antenna equations [10] which are further optimized to obtain better results. The optimized parameters of design model A are shown by Table 1.

2.2 Antenna Design B

For enhancing the gain of the antenna, the design has been modified and employed with H-shaped Defected Ground Structure which is created at the transmission line at the ground as shown in Fig. 2.

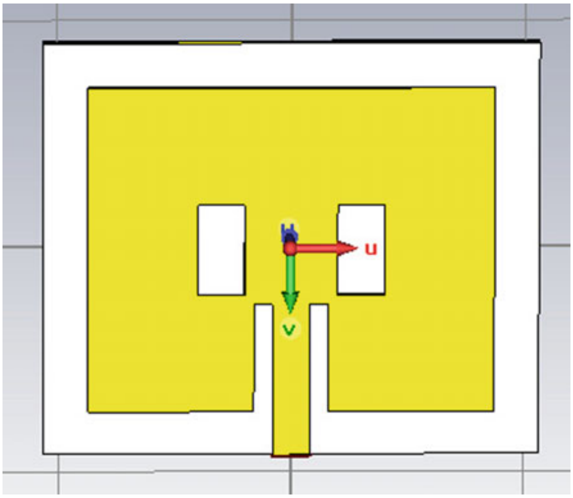


Fig. 1 Microstrip patch antenna without DGS

Table 1 Antenna A optimized parameters

Dimensions of	Length (mm)	Width (mm)
Main patch	71.6	87.2
Slot patch 1	20	10
Slot patch 2	20	10
Feed	35.41	4
Wave port	8	4.5

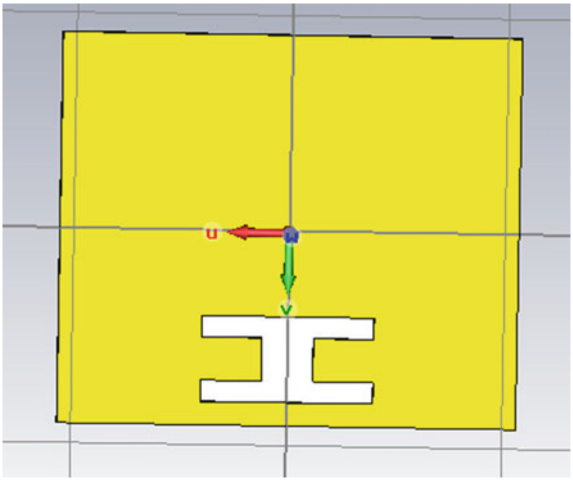


Fig. 2 Ground plane with H-shaped slot

The dimensions of the patch and slots of antenna are same as the previous design. An H-shape slot is further cut at the transmission line on the antenna ground whose measurements have been mentioned below in Table 2.

Table 2 Optimized parameters of antenna B

Dimension of	Length (mm)	Width (mm)
Main patch	71.6	87.2
Slot patch 1	20	10
Slot patch 2	20	10
Feed	35.41	4
Wave port	8	4.5
Ground H-slot		
Part 1	40	5
Part 2	10	8
Part 3	40	5

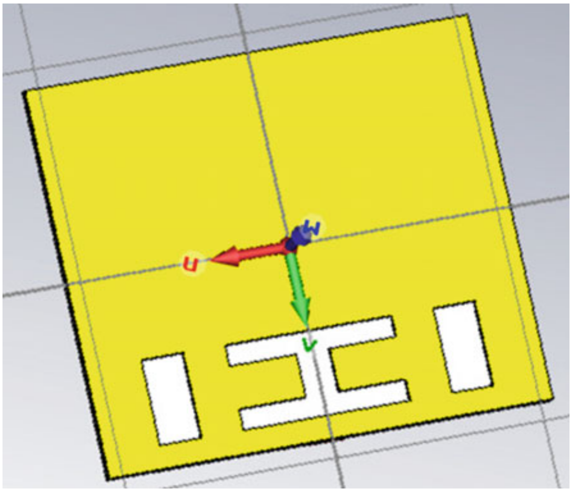


Fig. 3 Ground plane with H-shaped slot and two more slots

2.3 Antenna Design C

Figure 3 shows the design with DGS and extra two slots cut near the H-slot at the ground plane. This was done to improve the gain of the design. Gain improvement is seen in the results mentioned later.

The dimensions of the patch and H-shape slot are same as the previous design. Two more slots are further cut on the antenna ground whose measurements have been mentioned below in Table 3.

Table 3 Optimized parameters of antenna C

Dimension of	Length (mm)	Width (mm)
Main patch	71.6	87.2
Slot patch 1	20	10
Slot patch 2	20	10
Feed	35.41	4
Wave port	8	4.5
Ground H-slot		
Part 1	40	5
Part 2	10	8
Part 3	40	5
Ground slot 1	20	10
Ground slot 2	20	10

Table 4 Parameters of antenna for results and discussion

Design	Frequency	Return loss	Gain (dB)
A	2.42	−36.95	3.68
	0.90	−17.02	
B	2.16	−16.43	6.78
	1.44	−11.42	
C	2.03	−10.50	8.10
	1.80	−12.50	
	0.90	−13.50	

3 Results and Discussion

Three microstrip patch antennas with an inset feed have been designed. The first antenna Design A does not provide a good gain. The return loss is found to be good. The Design B was proposed provided with an H-shaped slot with a DGS and this resulted in a good gain. The results show that with inclusion of DGS the resonant frequency is shifted to the lower end because of the change in the current distribution of the radiating patch. The return loss results show that the antenna can be used for multiband. Another antenna Design C which is proposed increased the gain value and the return loss graph shows that the antenna can be used for multiband. Table 4 shows the comparison between the three antenna designs.

3.1 Results of Design A

(a) Return Loss

For the antenna without defect ground structure return loss is shown in Fig. 4. For the resonant frequency 2.45 GHz, a resonant peak (RL = −36.95 dB) for the patch antenna is obtained at 2.42 GHz which is below −10 dB.

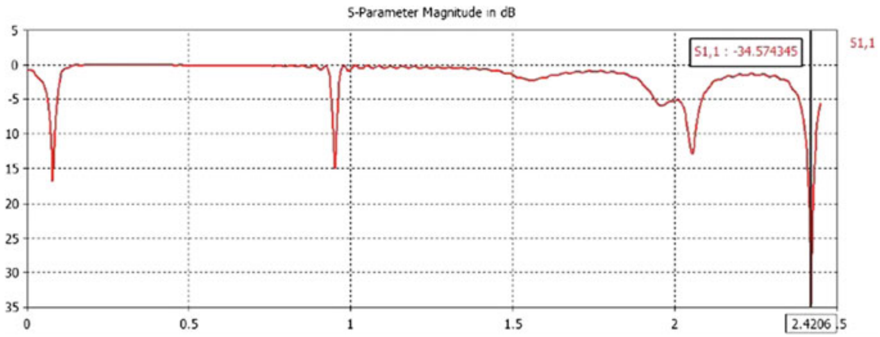
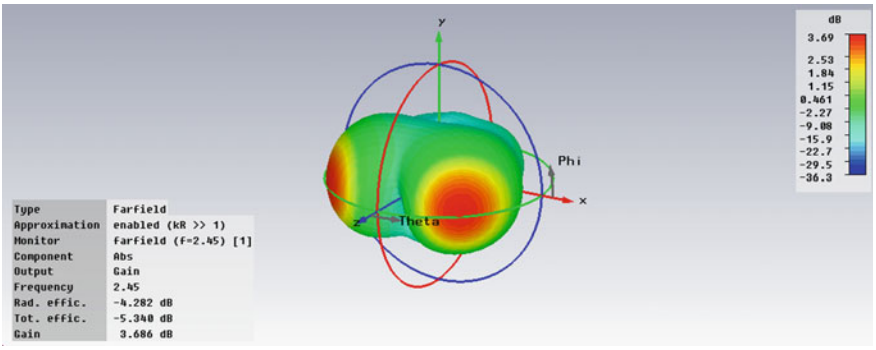


Fig. 4 S_{11} of the antenna A



Type	Farfield
Approximation	enabled (kR >> 1)
Monitor	farfield (f=2.45) [1]
Component	Abs
Output	Gain
Frequency	2.45
Rad. effic.	-4.282 dB
Tot. effic.	-5.340 dB
Gain	3.686 dB

Fig. 5 Radiation pattern showing gain in dB

(b) Radiation Pattern

The radiation pattern shown in Fig. 5 gives a gain of 3.68 dB.

3.2 Results of Design B

(a) Return Loss

For the patch antenna with H-shaped slot DGS is shown in Fig. 6 the return loss is obtained at 2.16 GHz which is below −10 dB. (Frequency shifted at lower end because of DGS).

(b) Radiation Pattern

The radiation pattern shown in Fig. 7 gives a gain of 6.788 dB.

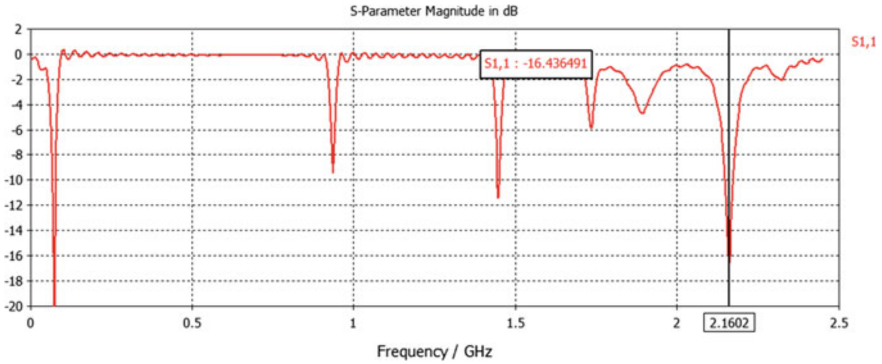
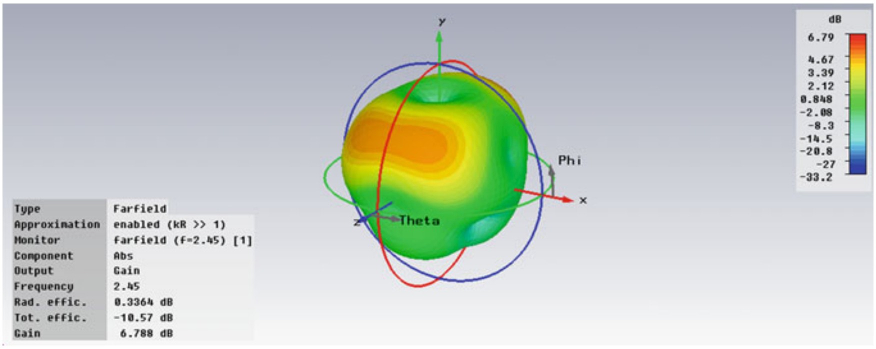


Fig. 6 S_{11} of the antenna B



```
Type            Farfield
Approximation   enabled (kR >> 1)
Monitor         farfield (f=2.45) [1]
Component       Abs
Output          Gain
Frequency       2.45
Rad. effic.     0.3364 dB
Tot. effic.     -10.57 dB
Gain             6.788 dB
```

Fig. 7 Radiation pattern showing gain in dB

3.3 Results of Design C

(a) Return Loss

For the patch antenna design C, return loss is obtained at 2.03 GHz which is below −10 dB given in Fig. 8.

(b) Radiation Pattern

The radiation pattern shown in Fig. 9 gives an enhanced gain of 8.10 dB.

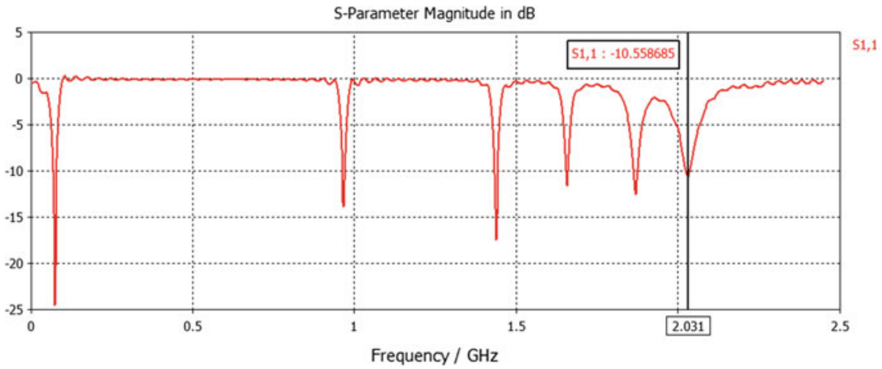
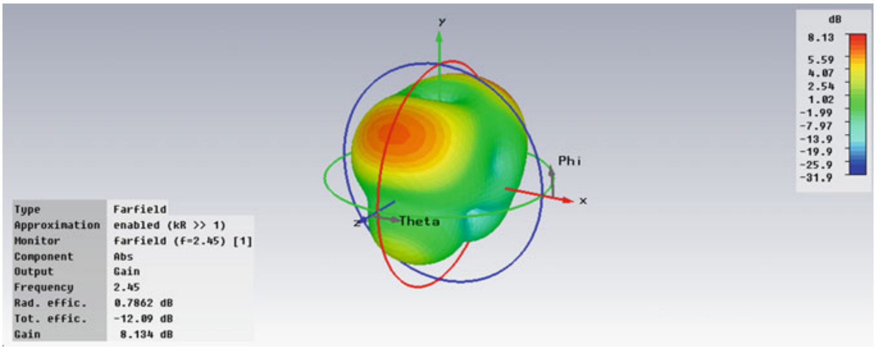


Fig. 8 S$_{11}$ of the antenna C



Type	Farfield
Approximation	enabled (kR >> 1)
Monitor	farfield (f=2.45) [1]
Component	Abs
Output	Gain
Frequency	2.45
Rad. effic.	0.7862 dB
Tot. effic.	-12.09 dB
Gain	8.134 dB

Fig. 9 Radiation pattern showing gain in dB

4 Conclusion

After analyzing the simulation results, we can see that with DGS the gain of the antenna is found to increase up to 6.78 dB in comparison to the antenna without DGS having a gain of 3.68 dB. Addition of one more DGS shows a further increase of gain up to 8.10 dB. Increase in gain because of DGS is attributed to the distribution of power to the fundamental frequency. Thus, there is 74% increase in gain when compared to the antenna without DGS. Multiband results show that the antenna with DGS can be used for IoT applications as well as for GSM.

References

1. M. Khayat, J. T. Williams, D. R. Jakson, and S. A. Long, “Mutual coupling between reduced surface-wave microstrip antennas,” IEEE Trans. on Antenna and Propagation, vol.48, Oct. (2000) 1581–1593
2. J. Park, A. Herchlein, and W. Wiesbeck, “A Photonic Bandgap (PBG) structure for guiding and suppressing surface waves in millimeter-wave antennas”, IEEE Transaction on Antennas Propagation, vol. 49, Oct. (2001) 1854–1857
3. R.G. Rojas, and K.W. Lee, “Surface wave control using nonperiodic parasitic strips in printed antennas”, IEE Pros.-Microwave, Antennas Propagation, vol. 148, Feb. (2001) 25–28
4. A.K. Arya, M.V. Kartikeyan, A. Patnaik, “Efficiency Enhancement of Microstrip Patch Antenna with Defected Ground Structure”, MICROWAVE, (2008) 729–731
5. Ashwini K. Arya, M.V. Kartikeyan, A. Patnaik, “Defected Ground Structure in the perspective of Microstrip antenna”, Frequenz, Vol. 64, Issues 5–6, Oct (2010) 79–84
6. A. K. Arya, A. Patnail, M.V. Kartikeyan, “Gain Enhancement of Micro-strip patch antenna using Dumbbell shaped Defected Ground Structure”, IJSRET, Vol. 2, Issue 4, July (2013) 184–188
7. EM simulator, CST Microwave studio TM, V.9
8. S. Bhunia, M.-K. Pain, S. Biswas, D. Sarkar, P. P. Sarkar, and B. Gupta, “Investigations on Micro strip Patch Antennas with Different slots and Feeding Points”, Microwave and Optical Technology Letters, No. 11, Vol. 50, November (2008) 2754–2758
9. S. Satthamsakul, N. Anantrasirichai, C. Benjangkaprasert and T. Wakabayashi, “Rectangular patch antenna with inset feed and modified ground plane for wideband antenna”, SICE Annual Conference 2008, The University Electro-Communications, Japan, August (2008) 20–22
10. Constantine A Balanis, Antenna Theory Analysis and Design, 2nd Edition, Singapore, John Wiley and Sons, (2002)

Robust Acoustic Echo Suppression in Modulation Domain

P.V. Muhammed Shifas, E.P. Jayakumar and P.S. Sathidevi

Abstract The presence of acoustic echo deteriorates the quality of speech transmission in mobile communication systems. In conventional acoustic echo suppression (AES) set-up, the echo path effect is modelled either in time domain or in frequency domain, and to cancel the echo, a replica of the echo is created by estimating the echo path response adaptively. Recently, the modulation domain analysis which captures the human perceptual properties is widely being used in speech processing. Modulation domain conveys the temporal variation of the acoustic magnitude spectra. In this work, a novel method for modelling the echo path and estimating the echo in modulation domain is developed and implemented. Echo cancellation is done effectively using the modulation spectral manipulation. So far, no work on echo suppression in modulation domain has been found as reported. The quality of output of the proposed system is found to be better than conventional AES systems.

Keywords Acoustic echoes · Echo path response · Modulation domain · Acoustic echo suppression

1 Introduction

The acoustic echoes in hands free devices arise due to the reverberation of the incoming far-end speech in the near-end environment, which will cause deterioration of the quality of speech transmission. Hence, it is essential to have a set-up that can effectively suppress this effect. There exist different acoustic echo cancellation (AEC) methods in literature for handling the echoes. In the conventional echo cancellation

P.V. Muhammed Shifas (✉) · E.P. Jayakumar · P.S. Sathidevi
Department of ECE, National Institute of Technology Calicut, Calicut, Kerala, India
e-mail: shifaspv001@gmail.com

E.P. Jayakumar
e-mail: jay@nitc.ac.in

P.S. Sathidevi
e-mail: sathi@nitc.ac.in

P.K. Sa et al. (eds.), *Progress in Intelligent Computing Techniques: Theory, Practice, and Applications*, Advances in Intelligent Systems
and Computing 719, DOI 10.1007/978-981-10-3376-6_57

the echo path is modelled as a filter with finite coefficients, and the filter coefficients are adaptively estimated using conventional adaptive algorithms [1]. An estimate of echo is computed and it is deducted from the microphone output.

In practice, the echo path cannot be modelled with finite coefficients since the echo path is constantly changing due to the changes in the surrounding environment. Thus the AEC will be effective only when the instantaneous changes are captured by the adaptive filter. This will result in the presence of little uncancelled echo in the signal that is transmitted back to the far-end side called as residual echo. Hence, in all practical AEC there will be a module to handle this unwanted residual echo.

Recently the echo cancellation by exploring the spectral manipulation is studied and shown that the acoustic echo suppression (AES) gives a robust performance when compared to AEC, especially under double-talk situations [2]. Since the human ear is insensitive to the phase variations of the speech, the spectral manipulation is bounded to the magnitude spectrum. In literature [3], Feller and Tournery had proposed a new method of estimating echo by extracting the echo path features, without estimating the complete echo path response. In this approach, the entire problem of echo estimation is reduced into the estimation of the delay introduced by the echo path and the coloration effect filter coefficient which captures the information about the spectral modification by the echo path. The coloration filter coefficients are estimated in acoustic domain as the correlation between the loud speaker and microphone signals. Since this modelling is almost insensitive to echo path changes, it will give robust performance with reduced computational complexity.

In the proposed method, AES is done in the modulation domain which captures the temporal variation of the acoustic magnitude spectrum, where the linguistic information is gathered. The echo path is modelled as a system which modifies the incoming far-end signals modulation spectrum. The detailed discussion about modulation domain analysis is given in the following sections. Section 2 explains modulation domain analysis in detail. The proposed modulation domain acoustic echo suppression (MDAES) is explained in Sect. 3. The implementation results and discussion are included in Sect. 4. Paper is concluded in Sect. 5.

2 Modulation Domain Analysis

Zadeh introduced a new dimension for speech analysis in the modulation domain [4]. He visualized the speech signal as a modulated signal, where the information carrying low frequency modulating signal modulates a high frequency carrier. Unlike the conventional modulation scheme, where the carrier frequency will be a pure tone, here the carrier frequency is composed of multiple frequencies. This concept ends up with a bi-frequency representation of the speech, which carries the linguistic information of the speech [5].

The extended popularity of the modulation domain analysis lies in the perceptual aspects of human auditory system. Bacon and Grantham [6] revealed the presence of channels in the auditory system and how these channels are tuned for the detection

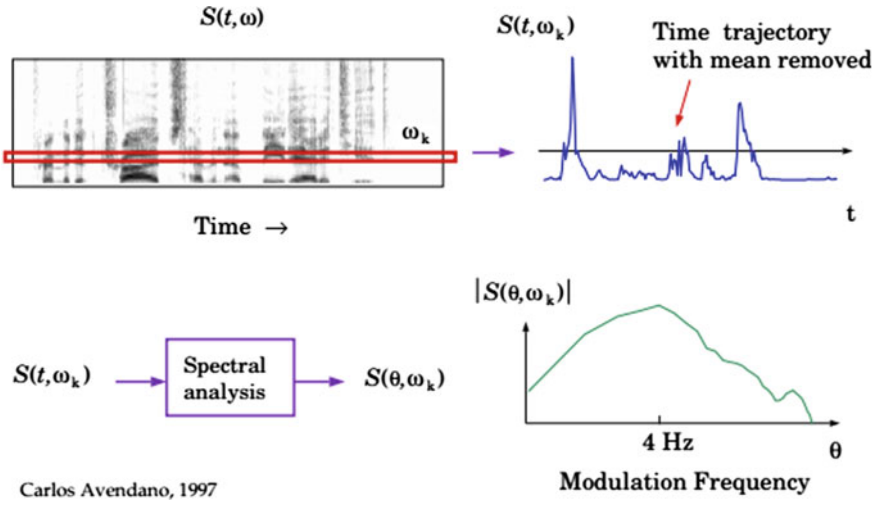


Fig. 1 Modulation frequency extraction

of modulation frequencies. In addition, the low frequency modulation sounds have been shown as fundamental information carrier in speech [7]. Drullman et al. [8, 9] have investigated the significance of modulation frequencies for the speech intelligibility by filtering the temporal envelopes of the acoustic frequency sub-bands. The analysis showed that the frequency components between 4 and 16 Hz are significant for intelligibility. The modulation spectrum measures the temporal variation of the vocal tract by taking the Fourier transform of the acoustic magnitude spectrum and becomes more robust towards the additive noises [10].

As specified in the above section, the modulation spectrum can be extracted through the spectral analysis of time trajectories of the power spectrum of the speech. The process of extracting the modulation spectrum from a speech segment is depicted in Fig. 1.

3 Modulation Domain Acoustic Echo Suppression (MDAES)

The modulation domain analysis of speech signal can be extended into the acoustic echo suppression module by exploring the modulation domain spectral manipulation techniques, which have been used in noise cancellation process and have shown better results as compared to the conventional frequency domain analysis with less annoying musical noises. In contrast to the conventional AES, where the echo path effect is modelled in frequency domain, we are modelling the effect in the modulation domain through the following formulations.

Let $x(n)$ and $y(n)$ be the loud speaker and microphone signals respectively and s(n) be the near-end speech. Let $x_d(n)$ is the delayed version of $x(n)$ after estimating the global delay parameter of the echo path. $X(k,fr,l)$, $Y(k,fr,l)$, $X_d(k,fr,l)$ and $S(k,fr,l)$ are the corresponding modulation spectra. Then, the input to the microphone will be,

$$Y(k,fr,l) = G_c(k,fr,l)X_d(k,fr,l) + S(k,fr,l) \tag{1}$$

where $G_c(k,fr,l)$ is the response of the echo path in modulation domain, the echo path modifies the modulation spectrum of the incoming speech. The symbols k, l and fr represent the acoustic frequency index, modulation frequency index and the frame index respectively.

Hence, the basic challenge of echo canceller is to estimate the gain effectively, create the replica of the echo and cancel it. If we can estimate the effect of echoes in the modulation spectra, we can effectively suppress these effects through the spectral modification as in the conventional noise suppression algorithms.

Multiply both sides of the Eq. (1) with $X_d{}^*(k,fr,l)$ and take the expectation. Since expectation is a linear operator, (1) becomes

$$\begin{aligned} E\{Y(k,fr,l)X_d{}^*(k,fr,l)\} = {} & E\{G_c(k,fr,l)X_d(k,fr,l)X_d{}^*(k,fr,l)\} \\ & +E\{S(k,fr,l)X_d{}^*(k,fr,l)\} \end{aligned} \tag{2}$$

Since the near-end speech $S(k,fr,l)$ and far-end speech $X_d(k,fr,l)$ are linearly independent to each other, the above expression becomes,

$$\begin{aligned} E\{Y(k,fr,l)X_d{}^*(k,fr,l)\} = {} & E\{G_c(k,fr,l)X_d(k,fr,l)X_d{}^*(k,fr,l)\} \\ & +E\{S(k,fr,l)\}E\{X_d{}^*(k,fr,l)\} \end{aligned} \tag{3}$$

Under the assumption that the speech signal having the distribution with zero mean and non-zero variance. i.e. $E\{S(k,fr,l)\} = 0$

$$E\{Y(k,fr,l)X_d{}^*(k,fr,l)\} = E\{G_c(k,fr,l)X_d(k,fr,l)X_d{}^*(k,fr,l)\} \tag{4}$$

Since the echo path response $G_c(k,fr,l)$ is a deterministic unknown, we can take it out from the expectation as

$$E\{Y(k,fr,l)X_d{}^*(k,fr,l)\} = G_c(k,fr,l)E\{X_d(k,fr,l)X_d{}^*(k,fr,l)\} \tag{5}$$

The final expression for the modulation domain coloration effect filter $G_c(k,fr,l)$ tracking the echo path will be a least square estimator as given in (6)

$$G_c(k,fr,l) = \frac{E\{Y(k,fr,l)X_d{}^*(k,fr,l)\}}{E\{X_d(k,fr,l)X_d{}^*(k,fr,l)\}} \tag{6}$$

Since the acoustic path changes continuously, it is better to estimate $G_c(k, fr, l)$ iteratively using the recursive relations with recursion parameter α,

$$\begin{aligned} E\{X_d{}^*(k, fr, l)Y(k, fr, l)\} = \alpha E\{X_d{}^*(k, fr-1, l)Y(k, fr-1, l)\} \\ +(1-\alpha)|X_d{}^*(k, fr, l)Y(k, fr, l)| \end{aligned} \tag{7}$$

$$\begin{aligned} E\{X_d{}^*(k, fr, l)X_d(k, fr, l)\} = \alpha E\{X_d{}^*(k, fr-1, l)X_d(k, fr-1, l)\} \\ +(1-\alpha)|X_d{}^*(k, fr, l)X_d(k, fr, l)| \end{aligned} \tag{8}$$

There exists two different situations in any echo cancellation set-up, namely single-talk and double-talk conditions. The single talk arises when there is no near-end speech such that the input to the microphone will only have the echo components whereas in the double-talk situation the microphone input is composed of both near end as well as the echo signals.

If we use the above least square estimator, the estimated filter coefficients would have been diverged from the actual value and will affect the quality of near-end speech transmission. To overcome this problem during the double talk, we should pause the filter adaptation during the double-talk situation. This is achieved by employing a double-talk detector (DTD) which detects the existence of double talk.

There exists different double-talk detection algorithms in both time and frequency domain [11, 12]. Since our system is operating in the modulation domain, we have proposed a correlation-based double-talk detector in the modulation domain by exploring the correlation between the signals, similar to the frequency domain method [11]. It is verified that the double-talk detector performs very well. The echo estimate in the modulation domain can easily be obtained by filtering the incoming far-end signal using the estimated coloration filter gain.

$$\widehat{Y}(k, fr, l) = G_c(k, fr, l)|X_d(k, fr, l)| \tag{9}$$

These estimates can be used for the spectral manipulations in the modulation domain spectral subtraction [13, 14]. Here the modification is being performed in the Fourier space of the spectral envelope whereas in the conventional frequency domain it is done in the magnitude spectral domain. The expression for the Wiener gain filter for the modulation spectral subtraction is

$$G(k, fr, l) = \left[\frac{max(|Y(k, fr, l)|^2 - \beta|\widehat{Y}(k, fr, l)|^2, 0)}{|Y(k, fr, l)|^2}\right]^{\frac{1}{2}} \tag{10}$$

where the parameter β represents the echo estimation efficiency. The parameter will be chosen a value greater than one, when echo is under estimated and less than one, when overestimated. As we are only considering the later reflections from the path while modelling the echo path, it is an under estimation case.

After calculating the gain filter, the final echo cancelled output will be obtained by filtering the microphone input using the estimated gain factor as,

$$E(k, fr, l) = G(k, fr, l)|Y(k, fr, l)| \tag{11}$$

The actual time domain expression of the echo cancelled signal that is transmitted back to the far-end side is obtained by taking the inverse transformations of the above expression

The perfection of echo cancellation depends on how well the echo component can be estimated from the microphone input. In conventional AES this separation is carried out in the usual Fourier domain, which captures the time variation of the signal whereas in this analysis in modulation domain we are measuring how quickly

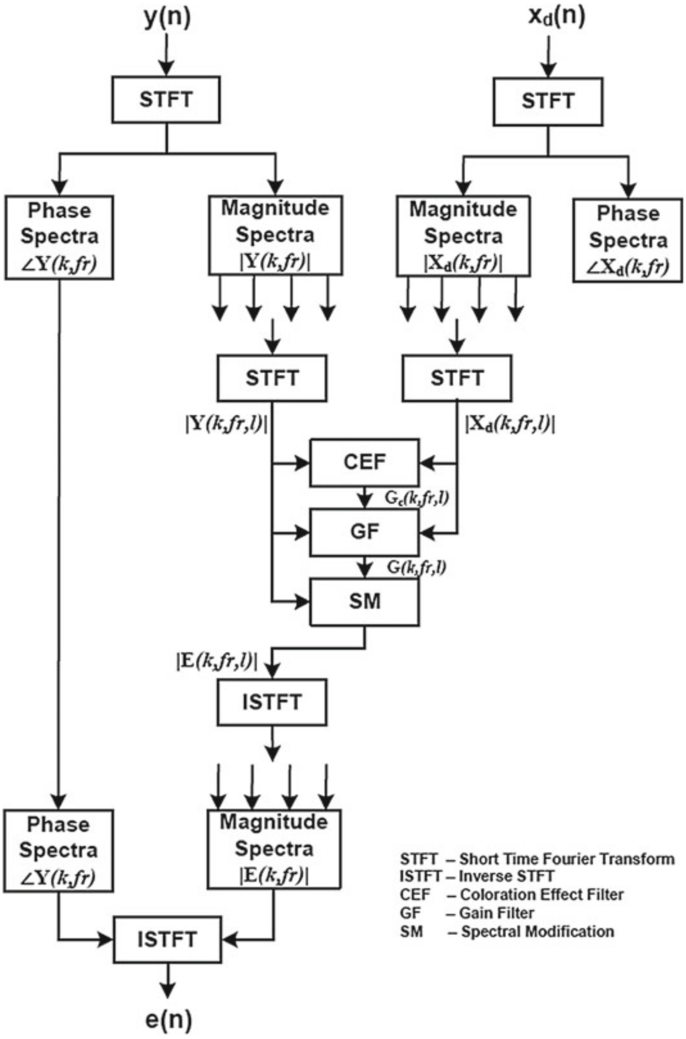


Fig. 2 Proposed MDAES algorithm

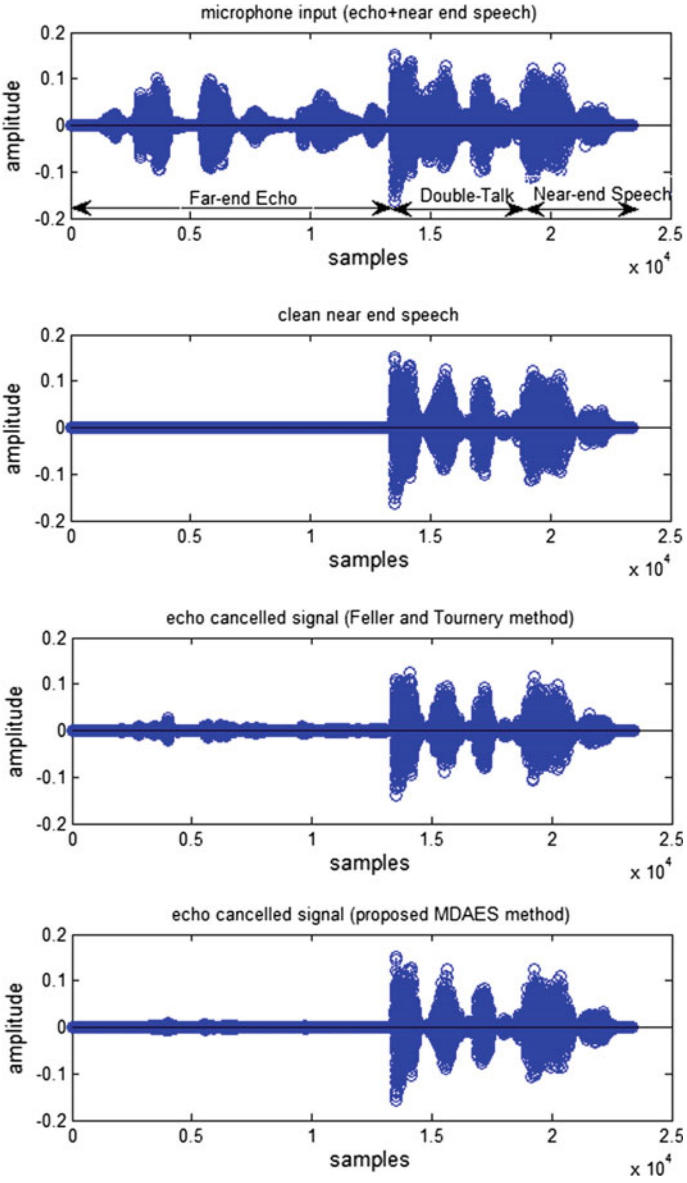


Fig. 3 Time domain representations of speech signals

the spectral components are varying across time. The block diagram of the proposed method is shown in Fig. 2.

The advantage of moving from the frequency domain to the modulation domain is evaluated in the next section. The performance of the proposed MDAES algorithm is compared with the existing frequency domain AES using objective as well as the subjective measures.

4 Experimental Analysis and Discussion

The evaluation of the proposed method is done on the utterance from the NOISEX data base with sampling frequency of 8 kHz [15]. The signal is processed as windowed frames of length 20 ms using hamming window with 50% overlapping. The 160 point FFT is performed to get into the acoustic domain. To get the modulation spectrum for a specific acoustic frequency, we took 160 point FFT of the consecutive acoustic spectral points of length 30 ms in time domain after windowed by hamming window of appropriate size.

The echo is created by filtering the far-end signal by a filter that models echo path to fit with a room with dimension $5 \times 4 \times 3\,\mathrm{m}^3$ with reflection coefficient 0.6

Fig. 4 ERLE plot of the above speech segment

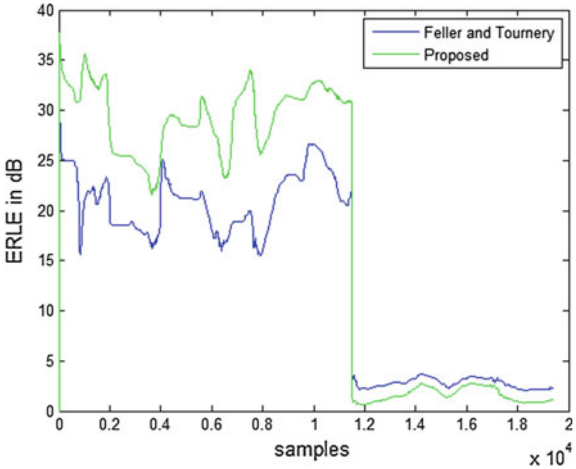


Table 1 Quality measurement of the processed speech

Quality measure	Feller and Tourney method	Proposed MDAES method
MOS	2.87	3.11
SA in dB	2.47	1.35

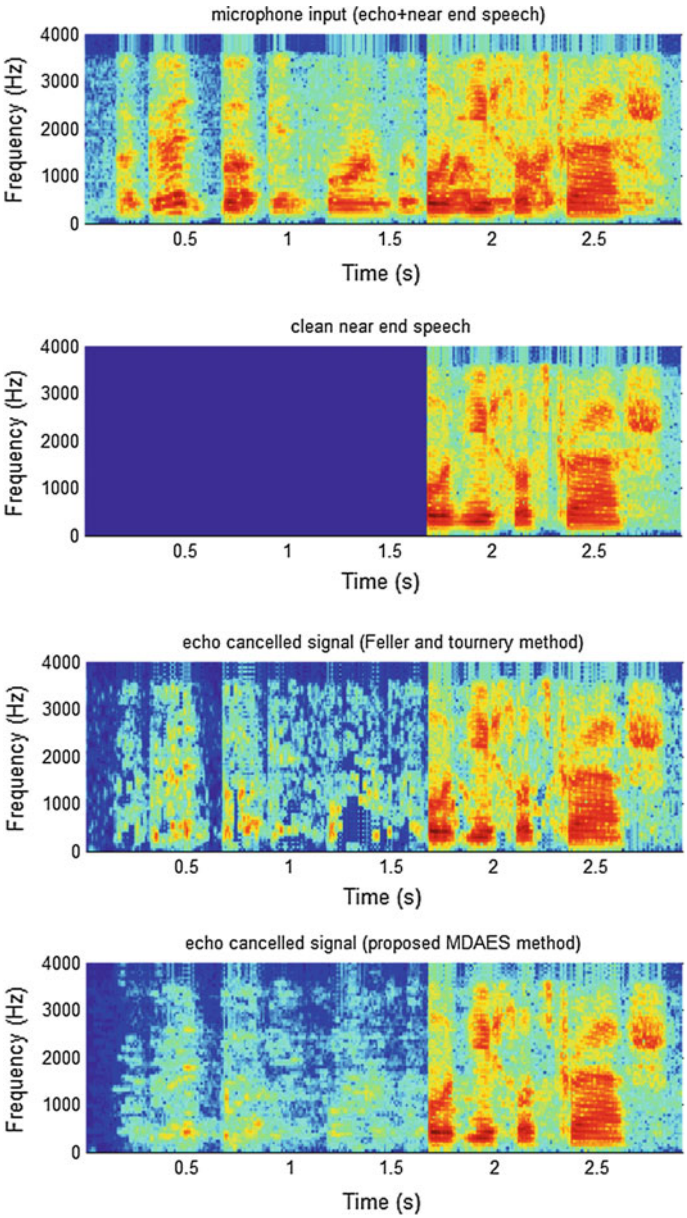


Fig. 5 Spectrogram of the above speech segments

and having 1400 coefficients [16]. The dialogue sequence starts with echo signal followed by the double-talk condition and ends with near end only utterance.

To compare the performance of the proposed method, we have implemented the AES set-up specified in [3] with the same parameters as specified above. The objective comparison is done by plotting the echo return loss enhancement (ERLE) [17].

The subjective comparison is done by calculating the mean opinion score (MOS) [18]. To calculate the MOS, we considered 10 dialogue sentences from the data base and calculated the average MOS. The Speech Attenuation (SA) of the near-end speech transmission is measured using the expression as in [17].

Further, to get visual differences between the two methods we have plotted the time domain and spectral domain features of the microphone input and clean speech along with the processed speech using both methods (Fig. 3).

From the above analysis, it is evident that the proposed MDAES method performs better as compared to the conventional method in all aspects. Figure 4 and Table 1 clearly indicate that the echo is suppressed efficiently in the proposed method during the single-talk case and gives less attenuation to the near-end speech signal during double-talk situations. Hence it acts as an efficient echo canceller. Further, the quality measurement table indicates the quality of the near-end speech processing through the echo cancellation set-up. It can be seen that, our method gives less attenuation to the near-end speech along with the improvement in perceptual quality (Fig. 5).

5 Conclusion

An efficient acoustic echo suppression system in modulation domain named as MDAES is developed and implemented in this paper. Both subjective and objective measures show the advantage of modulation domain analysis over the usual frequency domain echo cancellation. The performance of the proposed method is compared with that of the conventional AES algorithm. The proposed algorithm performs well in all aspects, providing better echo suppression without affecting the near-end speech transmissions. This is because, the modulation domain captures more information about the echoes than the conventional frequency domain and hence echo suppression is done very efficiently.

References

1. Sondhi MM: An adaptive echo canceler (1967) Bell Syst. Tech. J., vol. 46, 497–511.
2. Avendano C (2001) Acoustic echo suppression in the STFT domain. IEEE Workshop on Appl. of Sig. Proc.to Audio and Acoust: pp. 175–178.
3. Faller C and Tournery C (2006) Robust Acoustic Echo Control Using A Simple Echo Path. IEEE Int. Conf. Acoustic, speech, signal processing, (ICASSP): p. 12.
4. Zadeh, L. A. (1950). Frequency analysis of variable networks. Proceedings of the IRE, 38(3), 291–299.

5. Atlas L (2003) Modulation spectral transform: Application to speech separation and modification. Tech. Rep. 155, IEICE, Univ. Washington, Washington, WA, USA.
6. Bacon S, Grantham D (1989) Modulation Masking: Effect of modulation frequency depth and phase J. Acoust. Soc. Amer: pp. 2575–2580.
7. Atlas L, Shamma S (2003) Joint acoustic and modulation frequency. EURASIP J. Appl. Signal Process: pp. 668–675.
8. Drullman R, Festen J, Plomp R (1994) Effect of reducing low temporal modulations on speech reception. J. Acoust. Soc. Am. 95(5): pp. 2670–2680.
9. Drullman R, Festen J, Plomp R (1994) Effect of temporal envelope smearing on speech reception. J. Acoust. Soc. Am. 95(5): pp. 1053–1064.
10. Hermansky, H. (1998). Modulation spectrum in speech processing. In Signal Analysis and Prediction (pp. 395–406). Birkhuser Boston.
11. Das, Vineeta, Asutosh Kar, and Mahesh Chandra (2014) A new cross correlation based double talk detection algorithm for nonlinear acoustic echo cancellation. TENCON 2014–2014 IEEE Region 10 Conference: pp. 1–6.
12. Gnsler, T., Benesty, J. (2001). A frequency-domain double-talk detector based on a normalized cross-correlation vector. Signal Processing, 81(8), 1783–1787.
13. Paliwal, K., Wjcicki, K., Schwerin, B. (2010). Single-channel speech enhancement using spectral subtraction in the short-time modulation domain. Speech communication, 52(5), 450–475.
14. Togneri, R, Narasimha, M. (2015). Speech and Audio Processing for Coding, Enhancement and Recognition. T. Ogunfunmi (Ed.). Springer.
15. NOISEX-Speech at CMU. Rice University, USA. http://www.speech.cs.cmu.edu/comp.speech/noisex.html.
16. McGovern S, A Model for Room Acoustics, 2003. http://sgm-audio.com/research/rir/rir.html.
17. Lee, S. Y., Kim, N. S. (2007). A statistical model-based residual echo suppression. Signal Processing Letters, IEEE, 14(10), 758–761.
18. Loizou P: Loizou, P. C. (2013). Speech enhancement: theory and practice. CRC press.

FPGA-Based Equalizer Design Using a Novel Adaptive Reward-Punishment VSSLMS Algorithm for Rayleigh Fading Channel

Sudipta Bose and Iti Saha Misra

Abstract In this paper, a new and novel Reward-Punishment-based Variable Step Size Least Mean Square (RP-VSSLMS) algorithm has been proposed and a novel methodology is used to construct a Rayleigh fading channel adaptive equalizer employing the proposed algorithm in hardware domain. As the Rayleigh fading channel reveals the property of real-time wireless communication environment, it is chosen here. The Spartan 6 FPGA board is configured here to model the digital circuitry of the proposed RP-VSSLMS algorithm using a novel "Hardware Co-simulation" technique. The hardware co-simulation analysis showed that, the proposed RP-VSSLMS algorithm has faster convergence speed, smaller steady-state misadjustment, and lesser computational complexity than the existing LMS and VSSLMS algorithms. The performance of the proposed algorithm is observed by calculating the Bit Error Rate (BER) of different modulated signals under Rayleigh Fading channel.

Keywords Adaptive equalizer · VSSLMS · RP-VSSLMS · LMS · Rayleigh fading channel · Step size · BER · BPSK · QPSK · 8-QAM

1 Introduction

One of the most important problems in wireless communication is the execution of errorless detection and correction process at the receiving end. The wireless communication channel is time-varying in nature. Phenomenon like inter symbol

S. Bose (✉)
School of Mobile Computing & Communication,
Jadavpur University, Kolkata, India
e-mail: sudiptabose1991@gmail.com

I.S. Misra
Department of Electronics & Telecommunication Engineering,
Jadavpur University, Kolkata, India
e-mail: itisahamisra@yahoo.co.in

P.K. Sa et al. (eds.), *Progress in Intelligent Computing Techniques: Theory, Practice, and Applications*, Advances in Intelligent Systems and Computing 719, DOI 10.1007/978-981-10-3376-6_58

interference (ISI) is one of the largest obstacles in efficient transmission of data. It can be understood there is a need of establishment of reliable communication system eliminating the effect of interference at the receiver end. This concept leads to the formulation of channel equalization.

An adaptive equalizer can automatically adapt to the time-varying characteristics of the communication channel. This signal processing technique is used to counteract the effect of ISI [1]. To extract the error free desired signal from the corrupted signal at the receiver section, mechanism involving large-scale computation is required. Basically equalizer is classified into two types namely linear and nonlinear. As the linear equalizer has simple structure, it is used extensively in hardware-based implementation purpose of adaptive equalizer [2].

The Least Mean Square (LMS) Algorithm [3] has the ability to minimize error by adapting optimized tap weights in a tapped-delay-line equalizer. The weights of the equalizer are changed in response to the input sequence. The LMS algorithm is an example of adaptive signal processing. The LMS algorithm exhibits low complexity and it is easy to implement in hardware domain because of its simple structure. The importance of LMS algorithm is to update the tap weights recursively. The step size is the most important parameter of this algorithm. Large value of step size gives faster convergence speed and increased steady-state mean square error (MSE). In case of smaller value of the step size, the steady-state MSE is small, but the convergence rate is slow. It is possible to improve the performance of the LMS algorithm by making the value of step size variable instead of fixed. At the initial phase start with large step size for increased convergence rate and after reaching steady-state small step size is used to minimize the MSE. This can be implemented with variable step size LMS algorithm (VSSLMS). In [4], a VSSLMS algorithm has been discussed. This is known as Kwong's algorithm and it performs well in most of the operating conditions. But for the adjustment of the time-varying step size many parameters are needed which increases the computational complexity in presence of noise. A modified VSSLMS algorithm is thus proposed here, which is represented by Reward-Punishment based VSSLMS (RP-VSSLMS) algorithm and it improves the immunity of the algorithm under Rayleigh fading channel. The proposed algorithm gives great improvement in estimation speed and accuracy over LMS algorithm. Even its speed of convergence is comparable with Recursive Least Square (RLS) algorithm despite having lesser computational complexity than RLS algorithm.

It is evident from the literature survey that there is limited number of research works have been done on hardware-based implementation of adaptive equalizer. Here, the primary objective is the implementation of the newly proposed adaptive RP-VSSLMS algorithm in hardware platform considering real time wireless environment and performance analysis of the designed algorithm to be compared its performance with already existing adaptive algorithms. In Sect. 2, the proposed RP-VSSLMS algorithm used in this work is discussed. Section 3 includes the hardware co-simulation description. The performance analysis is represented in Sects. 4 and 5 gives the conclusion that we have derived from the results.

2 Proposed Algorithm

The basic flow of any VSSLMS algorithm closely follows the LMS algorithm described in [3, 5]. In case of communication through Rayleigh fading channel [6, 7], complex valued computation is required for the newly proposed RP-VSSLMS algorithm. Here an approach is made to derive a complex valued algorithm by extending the real-valued algorithm to allow complex signal processing that is widely used in real time wireless applications [8–10]. The input vector *X'(r)* and weight vector *W'(r)* are represented here,

$$X^{'}(r) = X_R(r) + iX_I(r) \tag{1}$$

$$W^{'}(r) = W_R(r) + iW_I(r), \tag{2}$$

where *R* and I represent the real and imaginary part of any complex valued signal respectively. The error response *e'(r)* and desired response *s'(r)* are represented by,

$$e^{'}(r) = e_R(r) + ie_I(r) \tag{3}$$

$$s^{'}(n) = s_R(r) + is_I(r) \tag{4}$$

Equalizer output

$$y^{'}(r) = w^{'T}(r)x^{'}(r) \tag{5}$$

Estimation error

$$e^{'}(r) = s^{'}(r) - y^{'}(r) \tag{6}$$

Tap weight adaptation

$$w^{'}(r+1) = w^{'}(r) + \mu(r)\,.\,e^{'}(r)x^{'}(r), \tag{7}$$

where *r* = number of iteration, *x'(n)* = input response at the iteration *r*, *w'(r)* = filter weight vector and *μ(r)* = the variable step size.

The required condition for stable operation is shown below,

$$0 < \mu(r) < 2\,/3\,tr\,[A] \tag{8}$$

where $A = E[x(r)\,x^T(r)]$, represents input autocorrelation matrix.

In this concept, if the result in one instance goes towards the converging value, then the value of adapting parameters are being "rewarded" so that in next iteration the result goes more near towards the converging value. Similarly, if the value moves away from the converging value, the adapting parameters are being "punished" so that in next iteration on the obtained resultant value restriction is imposed from the tendency of moving away from converging value. The definition of reward or punishment is generally subtraction or addition of a small positive valued Reward-Punishment coefficient (Δ). By employing this concept, the updating rule of the step size ($\mu(r)$) is as follows:

$$\text{If} \quad e^{'}(r) > e^{'}(r-1) \quad \mu(r+1) = \mu(r) + \Delta \tag{9}$$

$$\text{If} \quad e^{'}(r) < e^{'}(r-1) \quad \mu(r+1) = \mu(r) - \Delta \tag{10}$$

$$\text{If} \quad \mu(r+1) > \mu_{\max} \quad \mu(r+1) = \mu_{\max} \tag{11}$$

$$\text{If} \quad \mu(r+1) < \mu_{\min} \quad \mu(r+1) = \mu_{\min} \tag{12}$$

$$\text{Otherwise} \quad \mu(r+1) = \mu(r+1) \tag{13}$$

When *e'(r) < e'(r−1),* it is understood that the algorithm is moving towards convergence, and it is already known that when going towards convergence, the step size ($\mu(n)$) decreases. Thus the step-size adaptation process is being rewarded by decreasing the value by Δ. The mathematical operation is just reversed in case when *e'(r) > e'(r−1).* Then it is concluded that the algorithm is moving away from convergence. To restrict the dynamics, faster adaptation is required. Thus the step-size value is punished by increasing the value by Δ. $\mu(r)$ is bounded by upper value of μ_{max} and lower value by μ_{min}.

3 Designing FPGA Based System Model

To configure the FPGA board (Spartan-6 LX45) for implementing Rayleigh fading channel equalizer which employs the proposed RP-VSSLMS algorithm in hardware domain a new and novel approach known as “Hardware Co-simulation” is taken instead of writing direct VHDL code and then compiling the code by Xilinx compiler. In case of “Hardware Co-simulation”, combination of simulation-based platform MATLAB/SIMULINK and VHDL compiler Xilinx System Generator ISE design tool have been used. Here, Xilinx System Generator is a powerful tool that enables the use of SIMULINK platform for Hardware design.

For better understanding of the methodology of hardware co-simulation method an addition operation is shown in Fig. 1. This is a basic design to perform the addition operation between two integer values. The “Gateway IN/Gateway out” block is used to give input to the “Adder” (Xilinx-recognized special Simulink) block and getting the calculated result from it. Here, the “Adder” block generically contains the VHDL code to perform the addition operation. After compiling this program, the VHDL code is being converted to bit file that is the only recognizable entity by the hardware board. The results computed both in software and hardware domain is observed in the display.

Successful implementation of the proposed RP-VSSLMS algorithm for Rayleigh fading channel in FPGA platform is done by performing rigorous study and designing circuitry using Xilinx-Simulink Blocks avoiding the complexities of VHDL coding. The advantage of this methodology is that it makes the practical design simple, easy to implement, and cost-effective. The software–hardware interface makes it possible to programmatically use the hardware objects. The seamless interaction between the Simulink environment and FPGA board makes the simulation and hardware design much more convenient and provides results with high precision. Other advantages of hardware co-simulation are that it provides faster simulation, flexible modeling, and friendly graphical interface than normal software simulation, allowing the execution of large number of simulations with high accuracy. The disadvantage of this is that it has higher hardware resource utilization rate than HDL coding, but as the cost of additional hardware resources is comparatively very less so this can be ignored. The hardware co-simulation design of RP-VSSLMS algorithm consists of four major modules namely weight adaptation, tapped filter, add tree, and step-size adaptation module and it is represented by the following diagram shown in Fig. 2.

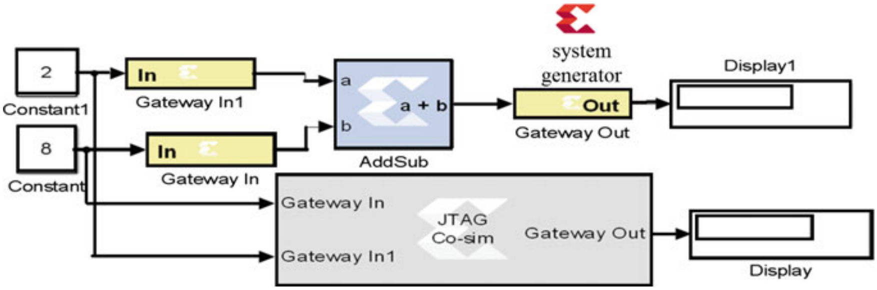


Fig. 1 An example of hardware co-simulation design of arithmetical operation

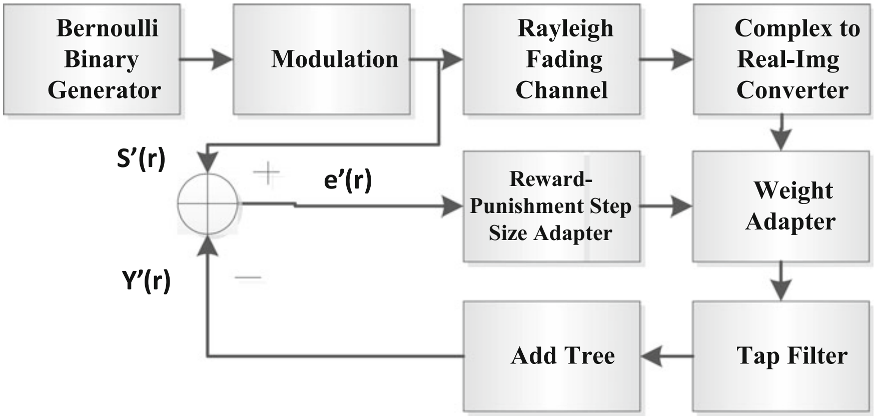

Bernoulli Binary Generator
Modulation
Rayleigh Fading Channel
Complex to Real-Img Converter
S'(r)
+
e'(r)
Reward-Punishment Step Size Adapter
Weight Adapter
Y'(r)
−
Add Tree
Tap Filter

Fig. 2 Steps of implementation of RP-VSSLMS algorithm under Rayleigh fading channel

The Bernoulli binary generator is utilized to generate random binary bit stream which acts as the training bits to configure the equalizer. The bit stream is then fed to a block representing modulation (BPSK, QPSK, 8-QAM) and then the modulated complex valued signal is fed through a Rayleigh fading channel with specific value of Standard deviation, Doppler shift, and SNR. Here, the modulated binary bit stream generation and its transmission through the Rayleigh fading channel is done completely in software domain using Simulink blocks. Only the execution of the RP-VSSLMS algorithm is done on the FPGA board. To perform this operation,

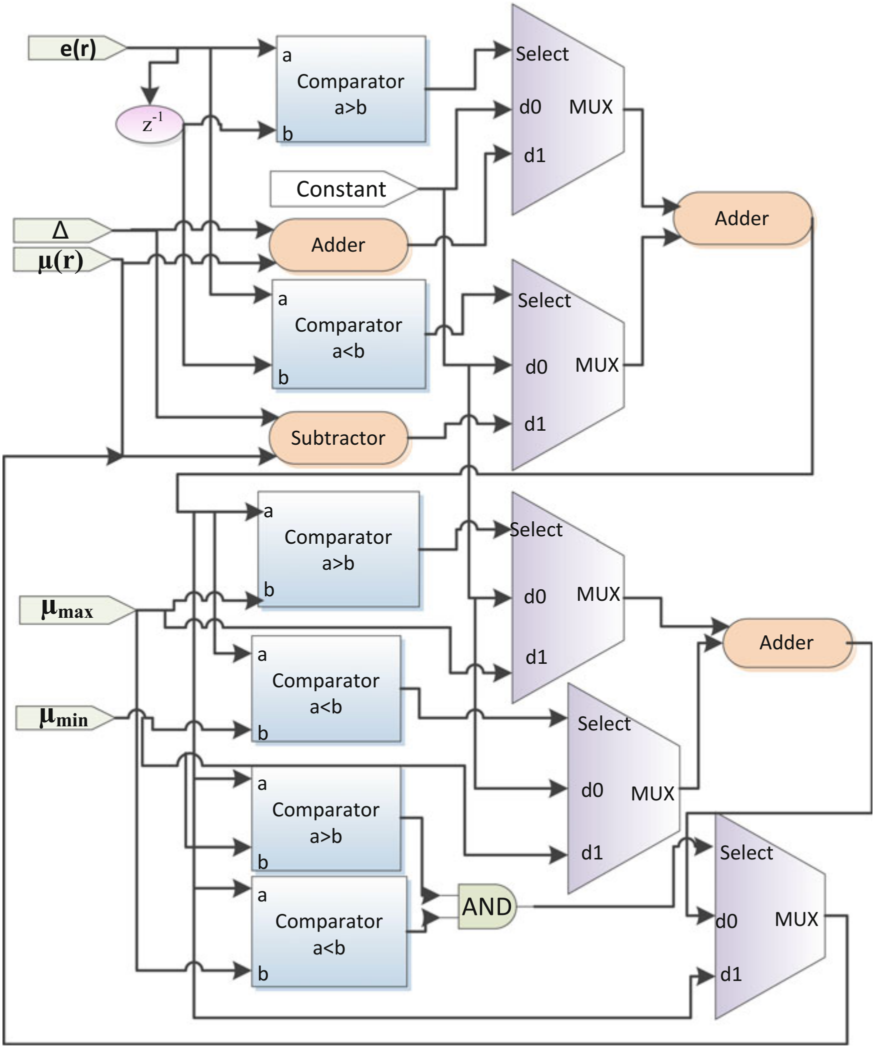


Fig. 3 Reward-punishment step size adapter module of RP-VSSLMS algorithm

the transmitted data is collected from the software section and the mathematical and logical calculations involved in the algorithm are done in hardware domain. The weight adapter module performs the tap-weight adaptation operation as stated in Eq. (7) and generates complex weight coefficients which depend on time-varying value of error. The tap filter module of 2 tap RP-VSSLMS executes the multiplication operation between the input and its delayed version with appropriate tap weight as stated in Eq. (5). The add tree module is used to perform the addition operation and calculate the actual output. The Reward-Punishment step size adapter module is the principle module that executes the mathematical and logical operations of the step size adaptation rules stated in Eqs. (9)–(13). Elementary blocks like multiplexer, comparator, and adder are used to implement this module in hardware domain as given in Fig. 3.

4 Results and Discussion

The performance of the implemented proposed RP-VSSLMS algorithm is evaluated by comparing it with the existing LMS algorithm and traditional VSSLMS algorithm under Rayleigh fading channel. The absolute error versus number of iteration has been plotted in Fig. 4 in order to examine the convergence performance. Using the Bernoulli binary generator 8 bits per symbol has been transmitted and the simulation is run for 1600 iterations. The convergence rate decreases as the Standard deviation (σ) increases as shown in Fig. 5, here σ varies from 0.316 to 0.707.

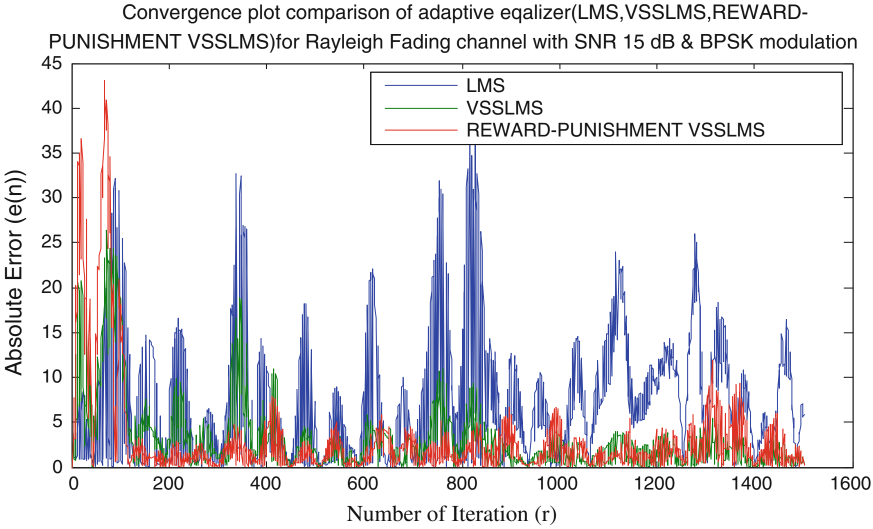


Fig. 4 Comparison of convergence rate of newly developed RP-VSSLMS algorithm and existing VSSLMS and LMS algorithm for Rayleigh fading channel (SNR 15 dB) with BPSK modulation

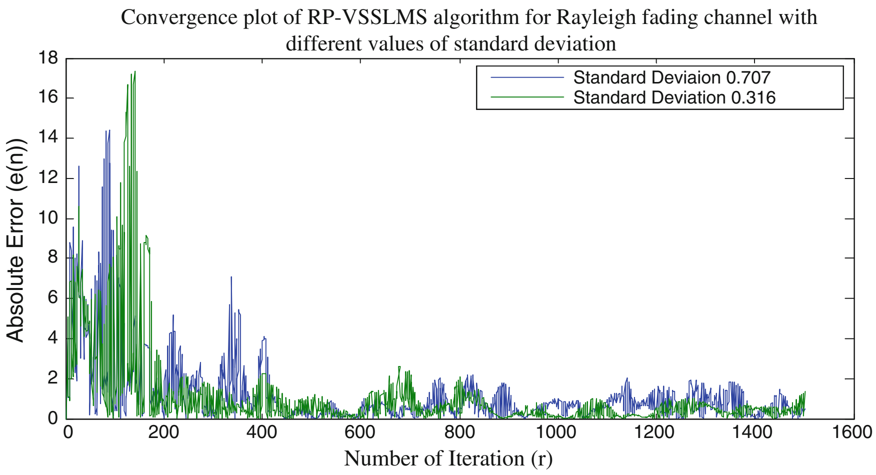


Fig. 5 Convergence plot of 2 tap RP-VSSLMS for Rayleigh fading channel with different values of standard deviation

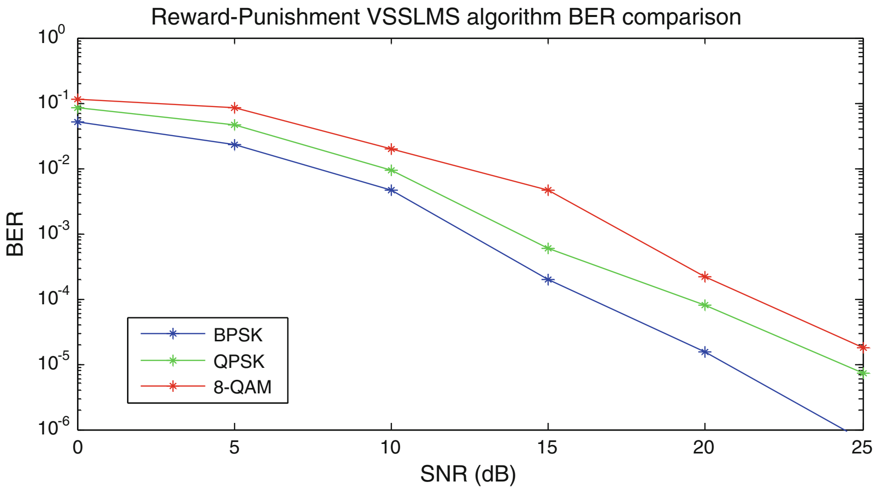


Fig. 6 BER versus SNR for (BPSK, QPSK, 8-QAM) signals using 2 tap RP-VSSLMS equalizer

The Bit error Rate (BER) is an important parameter for measurement of quality of the recovered data. In Fig. 6, the performance of the proposed algorithm is observed by measuring BER of different modulated signals like BPSK, QPSK, and 8-QAM under Rayleigh Fading channel from SNR 0 to 25 dB and it can be seen that for BPSK signal, the performance of RP-VSSLMS is better. In Fig. 7, the effect of the proposed RP-VSSLMS with BPSK modulation is compared with traditional LMS, VSSLMS algorithm in respect of BER. A comparative study has been performed on the hardware resource requirement of RP-VSSLMS and traditional

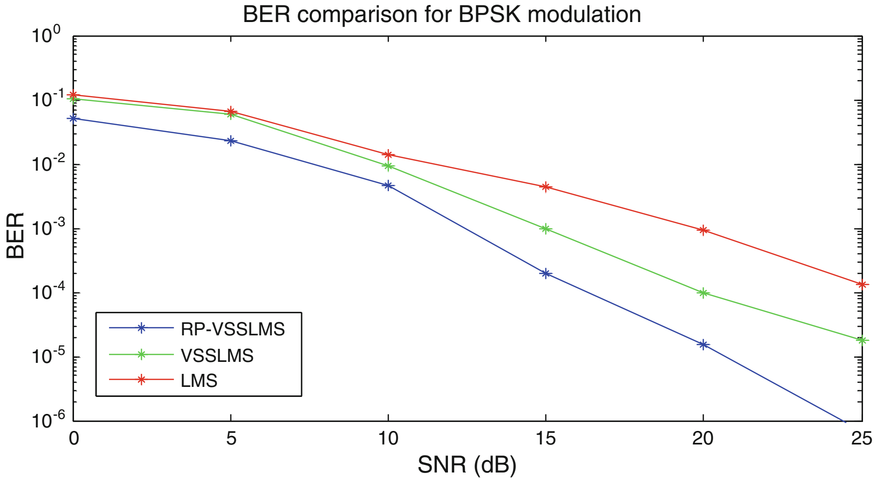


Fig. 7 BER versus SNR comparison of 2 tap RP-VSSLMS and existing LMS and VSSLMS algorithm under Rayleigh fading channel for BPSK modulation

VSSLMS which reveals that the resource requirement of the proposed algorithm is less than the traditional VSSLMS. It is therefore inferred that the proposed RP-VSSLMS algorithm is superior and gives much better performance in comparison to the existing LMS and VSSLMS algorithms.

4.1 Improved Performance Analysis of RP-VSSLMS Algorithm

The convergence rate of the proposed RP-VSSLMS algorithm for Rayleigh fading channel with same SNR value (15 dB) and modulation type (BPSK) is compared with the existing traditional LMS, VSSLMS to check the appropriateness of the error estimation process and improved performance of the implemented equalizer. Fixed value of step size 0.02 is used for the LMS algorithm and the parameter for RP-VSSLMS hardware co-simulation are used as, $\Delta = 0.008$, $\mu_{max} = 0.92$, $\mu_{min} = 0.002$.

The convergence of the newly developed 2 tap RP-VSSLMS algorithm is much faster than that the existing traditional 2 tap LMS and VSSLMS algorithm having less steady-state misadjustment. It can be observed that the error estimation process takes more than 1200 iterations for existing LMS algorithm and 800 iterations for existing VSSLMS algorithm to converge for Rayleigh fading channel having SNR 15 dB but for the proposed RP-VSSLMS channel it takes only 200 iterations to converge Fig. 4.

Convergence speed of RP-VSSLMS algorithm for Rayleigh fading channel has been measured for different values of standard deviation (σ) that is given in Fig. 5. It is seen that lower value of $\sigma = 0.316$ reveals faster convergence speed than higher value of $\sigma = 0.707$ because lower σ indicates the better channel condition and less channel variation with low Doppler shift (5 Hz), low delay spread (0.0795 µsec) and larger coherence bandwidth 2500 kHz than the higher σ, resulting the faster convergence.

The BER of BPSK, QPSK, and 8-QAM signals is calculated for Rayleigh fading channel at 10 Hz Doppler shift over 0 to 25 dB SNR as given in Fig. 6. With the increasing value of SNR, the BER will decrease in all of the three modulation techniques. Here, using BPSK for Rayleigh Fading better BER performance is obtained than both QPSK and 8-QAM. Similarly QPSK performs better than 8-QAM at higher SNR values. As, it can be seen from Fig. 6 that for BPSK modulation it takes nearly 13 dB SNR to maintain a BER of 10^{-3} but for QPSK it requires 18 dB SNR and 8-QAM requires 20 dB SNR to achieve a BER of 10^{-3}. Figure 7 reveals that in a Rayleigh fading channel, for BPSK modulation, the proposed RP-VSSLMS algorithm gives outstanding BER performance comparing with the existing LMS and VSSLMS algorithm where RP-VSSLMS needs 13 dB SNR but the existing VSSLMS needs 18 dB and LMS needs 24 dB SNR to maintain BER of 10^{-3}. The newly developed algorithm takes the performance to another level, outperforming the existing LMS, VSSLMS algorithm by giving approximately 28% better BER performance than VSSLMS and 45% better performance than LMS, thus standing as best among all other competitors in lower SNR condition.

Table 1 Hardware utilization data for 2 tap adaptive equalizer under Rayleigh fading channel

Parameter	Resources	Available	Utilized amount (2 tap VSSLMS algorithm)	Utilized amount (2 tap RP-VSSLMS algorithm)
Slice logic utilization	No. of slice registers	54576	1410 (1%)	1298 (1%)
	No. of slice LUT	27288	10808 (39%)	7806 (29%)
	No. of slice used as memory	6408	118 (1%)	110 (1%)
Slice logic distribution	No. of occupied slice	6822	1430 (20%)	1356 (20%)
	No. of LUT pair used		297	232
I/O utilization	No. of bonded I/O	218	0 (0%)	0 (0%)
	No. of LOCed I/O	1	0 (0%)	0 (0%)
Specific feature utilization	No. of RAMB16BWER	116	2 (2%)	2 (2%)
	No. of BUFG/BUFGMUXS	16	2 (16%)	2 (16%)
	No. BSCANS	4	1 (25%)	1 (25%)
	No. of DSP48AI	58	52 (89.65%)	41 (70.68%)
Peak memory			430 MB	390 MB

The analysis of the performance of the implemented equalizer from hardware resources utilization point of view is done here. In the Table 1, detailed information about the available resources on the Spartan 6 FPGA board and the amount utilized by the existing VSSLMS and the proposed RP-VSSLMS algorithm is given.

In case of Rayleigh fading channel as complex valued signal need to be processed so more resources are required for the implementation of 2 tap equalizer. As a result of less computational complexity than traditional VSSLMS algorithm, the proposed RP-VSSLMS algorithm requires less hardware resources than the traditional VSSLMS algorithm.

5 Conclusion

In this paper, a novel Reward-Punishment-based VSSLMS algorithm has been proposed and implemented for adaptive equalizer design. The hardware design of the 2 tap RP-VSSLMS algorithm is done successfully on the FPGA platform. The hardware implementation of adaptive equalizer is highly complex but it is necessary to design the equalizer considering real-time wireless environment. As a result Rayleigh fading channel has been chosen for the implementation of the proposed equalizer algorithm in the wireless communication system. The step-by-step design is implemented on FPGA using hardware co-simulation method. Here, 2 tap RP-VSSLMS has been implemented as the tap length could not be increased from 2 to further higher order due to hardware resource constraints. The advantage of hardware co-simulation is that it provides faster simulation with high accuracy than normal software simulation. Here, it can be seen that the implemented proposed 2-tap RP-VSSLMS offers faster convergence speed and smaller steady-state error compared to the traditional LMS and VSSLMS algorithm. It has less computational complexity and resource requirement than other available members of VSSLMS. It performs better at low SNR value in respect of BER performance and reveals drastic improvement in performance with respect to the existing LMS and VSSLMS algorithm. In further study the order of tap length can be increased to get a view about the change in performance. The advantage in the performance of error convergence rate and BER performance of the 2 tap RP-VSSLMS algorithm could be considered for use in wireless communications.

References

1. Schizas, ID., Mateos, G., Giannakis, GB.: Distributed LMS for Consensus-Based in-Network Adaptive Processing. IEEE Trans. Signal Process. 57(6) (2009) 2365–2382.
2. Widrow, B., McCool, J.M., Larimore, M., Johnson, C.R.: Stationary and non-stationary learning characteristics of the LMS adaptive filter. Proc IEEE. 64(8) (1976) 1151–1162.
3. Haykin, S.: Adaptive Filter Theory. 4th edn. Prentice-Hall, Englewood Cliffs (2000).

4. Kwong, RH., Johnston, EW.: A variable Step Size LMS Algorithm. IEEE Trans. Signal Process. 40(7) (1992) 1633–1642.
5. Bose, S., Mondal, A., Misra, IS.: FPGA Based Hardware Implementation of Adaptive Equalizer for Rayleigh Fading Channel. In Proc. of the ICMOCE, Bhubaneswar (2015) 338–341.
6. Lapidoth, A., Shamai (Shitz), S.: Fading channels: How Perfect Need Perfect Side Information be?. IEEE Trans. Inf. Theory. 48 (2002) 1118–1133.
7. Baddour, K.E., Beaulieu, N.C.: Autoregressive Modeling for Fading Channel Simulation. IEEE Trans. Wireless Commun. 4(4) (2005) 1650–1662.
8. Aboulnasr, T., Mayyas, K.: A Robust Variable Step-Size LMS-Type Algorithm: Analysis and Simulations. IEEE Trans. Signal Process. 45(3) (1997) 631–639.
9. Costa, MH., Bermudez, JCM.: A Robust Variable Step-Size Algorithm for LMS Adaptive Filters. In Proc. of the ICASSP, Toulouse (2006) 93–96.
10. Kang, R.H., Johnstone, E.W.: A variable Step Size LMS algorithm. IEEE Trans. Signal Process. 40(7) (1992) 1633–1642.

Phase Reversal and Suppressed Carrier Characteristics of Neo-Cortical Electroencephalography Signals

Manikumar Tellamekala and Shaik Mohammad Rafi

Abstract Transmission networks of human nervous system carry both sensory and motor neural signals simultaneously. For instance, median nerve carries sensory information from middle, index, and thumb finger to primary motor cortex. In communication engineering systems double-sideband suppressed carrier (DSB SC) modulation is a prominent power efficient analog modulation technique. This paper sheds some light on investigating the modulation technique that is followed by neural networks of human nervous system. EEG signals from motor cortex are applied as inputs to a narrow band pass filter with bandwidth 0.4 Hz. Characteristics of output signals from the filter are similar to that of amplitude modulated signals. In time domain, a 180° phase reversal is observed in case of all outputs. Suppression of power at a particular frequency and boosting the powers of frequencies which are at equidistant from the suppressed frequency value is the key feature of the filters output signal. Results of this study are prima facie evidences that allow one to think in the direction, human nervous system might be following double-sideband suppressed carrier modulation which is the most power efficient technique among available analog modulations, to avoid spectral overlapping.

Keywords Double-sideband suppressed carrier modulation • Phase reversal • Broca's area • Neo-cortices

M. Tellamekala (✉) · S.M. Rafi
Department of Electronics and Communication Engineering, Rajiv Gandhi University
of Knowledge Technologies, R.K. Valley, Nuzvid, India
e-mail: manikumarrkv@gmail.com

S.M. Rafi
e-mail: rafi@rguktrkv.ac.in

P.K. Sa et al. (eds.), *Progress in Intelligent Computing Techniques: Theory, Practice, and Applications*, Advances in Intelligent Systems
and Computing 719, DOI 10.1007/978-981-10-3376-6_59

1 Introduction

Understanding the functioning of human brain is the most essential aspect to address key challenges such as design and development of diagnostic and treatment tools for people with severe neurological disorders, implementation of self-assisting equipment for locked-in people, advancement of human–machine interaction mechanisms, etc. All over the world, about forty million people are suffering from Alzheimer's disease. This number keeps on increasing in recent years. Several such severe neural disorders are throwing brain teasing problems towards scientific and engineering communities. Treatment at early stages of this disease is very effective and critical to cure.

Electroencephalography became a popular brain imaging technology, for it offers a great temporal resolution in studying electrical patterns of different brain regions. Researchers have already demonstrated how one can employ modulation content in electroencephalography (EEG) signals, (strength of modulation (SOM) and phase of modulation (POM)), to develop semi-automated diagnosis of Alzheimer's disease [1, 2]. As shown in Fig. 1, sensory signals from index, middle fingers, and thumb travel to primary motor cortex through median nerve [3–5]. Likewise Ulnar nerve is responsible for carrying sensory signals from little and middle fingers [6–8]. What if both the little and middle fingers sensory signals have common frequency components with significant amount of information? Simultaneous traversal of two message signals with common frequency components through a single channel, results in spectrum overlapping or interference [9, 10]. Subsequently the entire communication process ends up with a huge loss of information. Essence of frequency division multiplexing (FDM) comes at this point

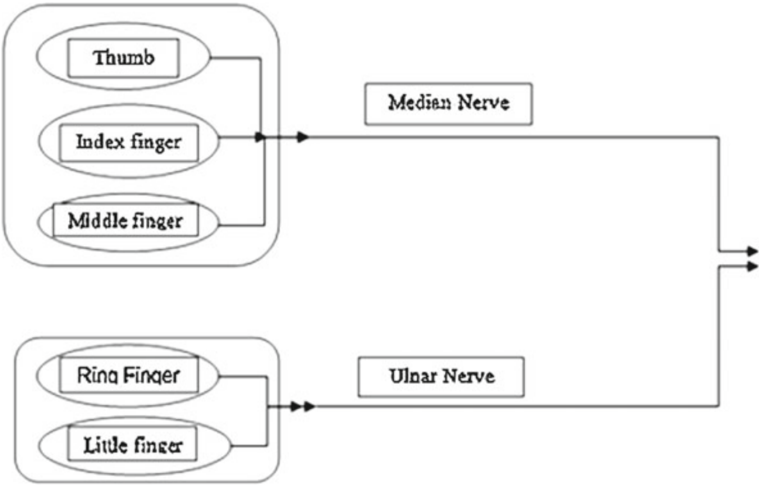


Fig. 1 Schematic block level diagram of neural networks of hand

[11–13]. In order to avoid the spectrum overlapping, one can shift the spectra of two message signals to different frequency positions on spectrum. Thus interference can be overcome without losing information.

2 Materials and Methods

To investigate the modulation technique that is being followed by biological neural communication networks, noninvasive EEG is used to collect signals from primary motor cortex. By using 10–20 international EEG electrode placement system, two active EEG electrodes are positioned at C3 and C4 points. A 50 Hz notch filter is employed to remove power line noise from EEG signals. Subjects are instructed to make two different movements with their left thumbs and left index fingers simultaneously.

Right hemispheres primary motor cortex transmits electrical patterns through the median nerve to generate intended motions with left thumb and left index finger. This is the same case with left hemispheres primary motor cortex and right thumb and right index finger.

In general, during motor activities mu rhythm (8–15 Hz) frequency band of EEG signals is dominant. EEG signals that are collected from both left and right hemispheres are passed through a Butterworth band pass filter of order 5 with bandwidth 0.5 Hz. Band limits of this filter are varied in between 8 and 15 Hz for every 0.4 Hz.

Output signal from the filter in time domain is shown in Fig. 2. This output signal looks like an amplitude modulated pattern. A 180° phase reversal can also be observed in this signal. Power spectrum of this output signal, which is also shown in Fig. 2, resembles that of double-sideband suppressed carrier modulation. Block level representation of a DSB SC modulation system and frequency domain characteristics of a DSB SC signal are displayed in Fig. 3.

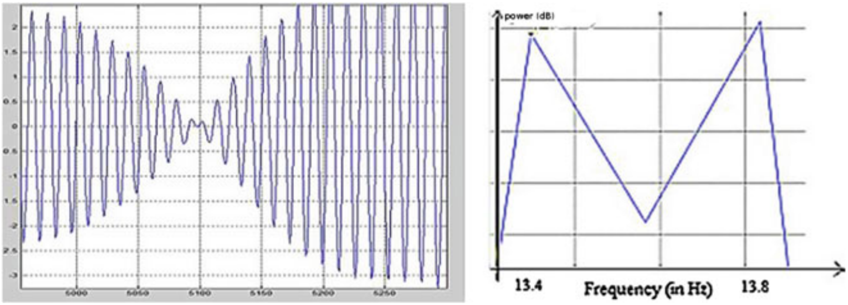


Fig. 2 Output from a band pass filter with band limits 13.4 and 13.8 Hz when EEG signal from motor cortex is applied as input and power spectrum of the filter's output

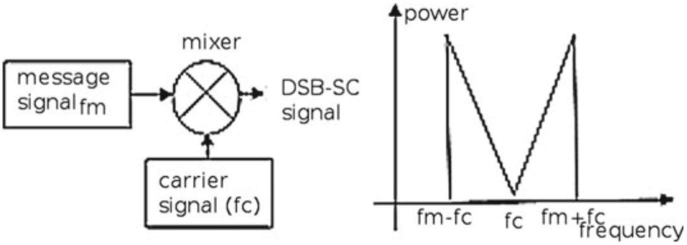
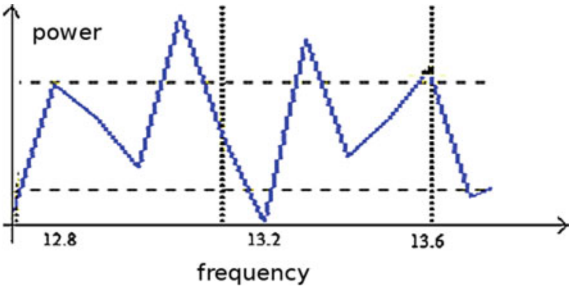


Fig. 3 Block diagram of DSB SC modulation and power spectrum of a DSB SC modulated signal

Fig. 4 Power spectrum of output from filter with band limits 12.8 and 13.6 Hz



Now, the bandwidth of the filter is increased from 0.4 to 0.8 Hz. Spectrum of the filters output for the same input EEG signal is displayed in Fig. 4. From this spectrum, one can primarily infer that two sinusoids with different frequencies are modulated with a single carrier signal whose power is suppressed.

Output signal from the filter when BCI IV competitions primary motor cortices EEG signal is applied as input and corresponding frequency spectrum are as shown in Fig. 5. In this case also amplitude modulation patterns, 180° phase reversals [14] are observed in time domain and carrier power suppression is observed in frequency domain. Experiments are performed on the signals that are collected from Broca's area, a language center in brain that is responsible for the production of speech and from primary visual cortex region. The envelope of the modulated signal is extracted from Hilbert's envelope extraction method. All typical subbands (Delta, Theta, Alpha, Beta, Gamma) and their corresponding envelopes (shown in red color) of the signal collected from Broca's area are shown in Fig. 6a. One of the subbands of the signal and the signal's Hilbert envelope are analyzed. Here Alpha band is shown in Fig. 6b, the 180° phase shift is observed in time domain. The spectrogram of the signal and the Hilbert envelope alone is plotted. Here the spectral characteristics clearly indicate the feature of modulation, the shift of the modulating signal, the Hilbert envelope, to the carrier frequencies. Hence same modulation phenomenon is observed in both time and frequency domains in the Broca's and Visual cortex regions also.

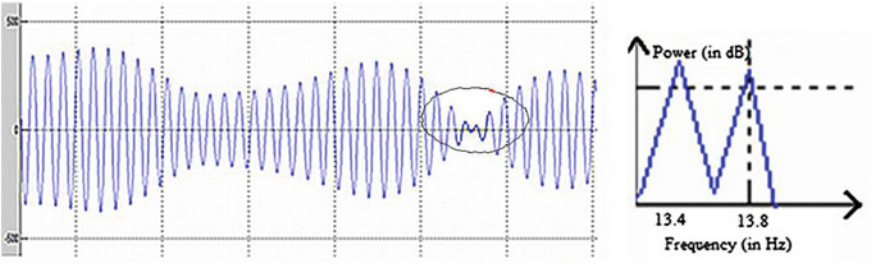


Fig. 5 Output from a band pass filter with band limits 13.4 and 13.8 Hz when motor EEG signal of BCI IV competition is applied as input and corresponding power spectrum of the signal

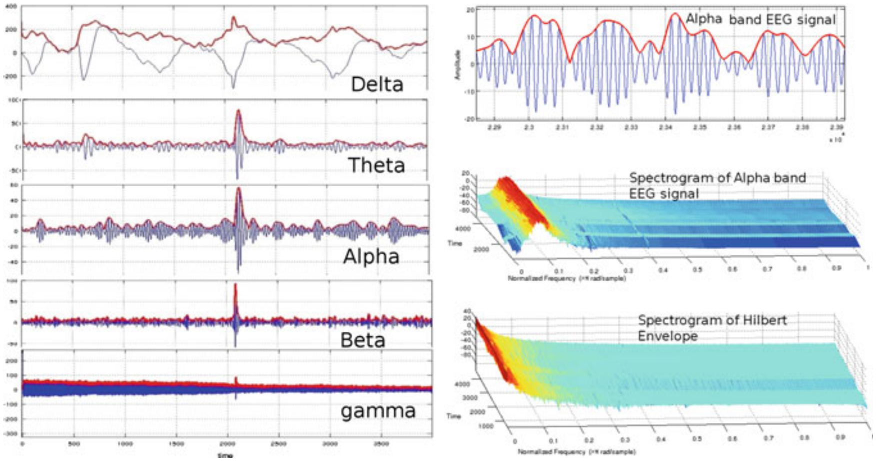


Fig. 6 Subbands and their Hilbert envelopes of EEG signal and alpha band (8–12 Hz signal) and spectrogram

3 Results and Analysis

After output signals from the filter, both in time and frequency domains, are insinuating that the modulation technique that is being followed by communication networks of human nervous system might be DSB SC modulation. As the carrier is suppressed, DSB SC is the most power-efficient modulation mechanism among all the available analog modulations. As nervous system consumes huge amount of energy for its functioning, in order to optimize power consumption, evolution of mammalians brain architecture might have chosen DSB SC modulation, for it offers great power efficiency.

A clear understanding of modulation mechanism that is associated with biological neural networks unravels the underlying mechanisms of brain teasers in neuroscience. For instance in bi manual interference two same geometric shapes can be drawn simultaneously by using left and right hand but drawing two different geometrical shapes, is highly impossible.

Accurately deciphering and classifying EEG signals of different cognitive activities under different environmental conditions is a challenging task. If communication networks in human nervous system are following modulation to do frequency multiplexing of message signals, some groups of neurons might be acting as blocks of a DSB SC modulator, as shown in Fig. 3, at transmitter side. At receiver side also demodulating blocks might present. Subsequently frequency range of a message signal purely depends upon the available spectrum after allocating carrier frequencies for other message signals in that same channel. Future steps in this direction may make us more knowledgeable about communication architecture of human brain. This would be a helpful solution to overcome current limitations of diagnostic and treatment tools of neurological disorders, brain computer interfaces and real genuine artificial intelligence.

If neural systems have adopted DSB SC modulation then the immediate question is about the process of demodulation. Phase locked loop, comprises of a multiplier, low pass filter and a voltage controlled oscillator blocks, is a key system in a squaring loop. In neo-cortical regions of human brain, it is observed that phase locking phenomenon is taking place and this is the basis for Steady State Evoked Potentials (SSVEP) experiments. Using squaring loop is one of the demodulation techniques of DSB SC signals. Though computational complexity is high in the process of DSB SC demodulation, nervous system might be using DSB SC because of low power consumption.

Further studies in this direction may provide deeper insights into the communication network architectures of neural networks and subsequently significant improvement in the performance of neural engineering applications. There is a possibility that demodulated EEG signals may boost up classification accuracy rate in Brain computer Interfaces. Frequency modeling of channel characteristics can be done based on the discrepancy between spectra of two side bands of same message frequency. This model may reduce the performance gap between noninvasive and invasive EEG electrodes.

References

1. F. J. Fraga, T. H. Falk, P. A. M. Kanda, and R. Anghinah, "Characterizing Alzheimer's Disease Severity via Resting-Awake EEG Amplitude Modulation Analysis," PLoS One, vol. 8, no. 8, 2013.
2. L. R. Trambaiolli, T. H. Falk, F. J. Fraga, R. Anghinah, and A. C.Lorena, "EEG spectro-temporal modulation energy: A new feature for automated diagnosis of Alzheimer's disease," in Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS, 2011, pp. 38283831.
3. M. L. Boninger, B. G. Impink, R. A. Cooper, and A. M. Koontz, "Relation between median and ulnar nerve function and wrist kinematics during wheelchair propulsion," Arch. Phys. Med. Rehabil., vol. 85, no. 7, p. 11411145, 2004.

4. J. M. Brown, A. Yee, and S. E. MacKinnon, “Distal median to ulnar nerve transfers to restore ulnar motor and sensory function within the hand: Technical nuances,” Neurosurgery, vol. 65, no. 5, pp. 966977, 2009.
5. J. B. Jaquet, A. J. Luijsterburg, S. Kalmijn, P. D. Kuypers, A. Hofman, and S. E. Hovius, “Median, ulnar, and combined median-ulnar nerve injuries: functional outcome and return to productivity.,” J. Trauma, vol. 51, no. 4, pp. 687692, 2001.
6. R. P. Calfee, P. R. Manske, R. H. Gelberman, M. O. Van Steyn, J. Steffen, and C. A. Goldfarb, “Clinical assessment of the ulnar nerve at the elbow: reliability of instability testing and the association of hypermobility with clinical symptoms.,” J. Bone Joint Surg. Am., vol.92, no. 17, pp. 28012808, 2010.
7. P. Rajan, R. Premkumar, P. Rajkumar, and J. Richard, “The impact of hand dominance and ulnar and median nerve impairment on strength and basic daily activities,” J. Hand Ther., vol. 18, no. 1, pp. 4045, 2005.
8. D. B. Polatsch, C. P. Melone, S. Beldner, and A. Incorvaia, “Ulnar Nerve Anatomy,” Hand Clinics, vol. 23, no. 3. pp. 283289, 2007.
9. M. Loukas, R. G. Louis, L. Stewart, B. Hallner, T. DeLuca, W. Morgan,R. Shah, and J. Mlejnek, “The surgical anatomy of ulnar and median nerve communications in the palmar surface of the hand.,” J. Neurosurg.,vol. 106, no. 5, pp. 887893, 2007.
10. M. M. Hoogbergen and J. M. Kauer, “An unusual ulnar nerve-median nerve communicating branch.,” J. Anat., vol. 181 Pt 3, pp. 513516, 1992.
11. S. Slobounov, M. Hallett, C. Cao, and K. Newell, “Modulation of cortical activity as a result of voluntary postural sway direction: An EEG study,” Neurosci. Lett., vol. 442, no. 3, pp. 309313, 2008.
12. B.-K. Min and H.-J. Park, “Task-related modulation of anterior theta and posterior alpha EEG reflects top-down preparation.” BMC Neurosci., vol. 11, p. 79, 2010.
13. I. Choi, S. Rajaram, L. a Varghese, and B. G. Shinn-Cunningham, “Quantifying attentional modulation of auditory-evoked cortical responses from single-trial electroencephalography.” Front. Hum. Neurosci., vol. 7, no. April, p. 115, 2013.
14. B. Kanmani, “The phase-reversal in DSB-SC: a comment,” IEEE 13th Digital Signal Processing Workshop and 5th IEEE Signal Processing Education Workshop, DSP/SPE 2009, Proceedings, 2009, pp. 786790.

Short-Range Frequency-Modulated Continuous Wave (FMCW) Radar Using Universal Software-Defined Radio Peripheral (USRP)

Munesh Singh, Sourav Kumar Bhoi and Pabitra Mohan Khilar

Abstract In this paper, we design a prototype FMCW (Frequency Modulated-Continuous Wave) radar for short distance using universal software defined radio peripheral (USRP). USRP is an ideal platform for all type of telecommunication research, until the bandwidth requirement is fulfilled by universal serial bus (USB). The limited bandwidth of USB bus restricted to realize short distance radar on this platform using GNU-RADIO signal processing block. To avoid the bandwidth limitation, we build transmit section independently inside the FPGA, which free the transmit section bandwidth to be fully utilized by receiving section. The prototype performance test is carried out at the center frequency of 2.4 GHz with a bandwidth of 15 MHz. To avoid the clutter of the environment, directional antenna is used for better target detection at certain distances. The return target echo is further processed using MATLAB.

Keywords Gnu radio · USRP · Software define radar · DDS · Fmcw radar

1 Introduction

Radar is used for remote sensing of target activity such as its speed, direction, and location [1], [2], [3]. In this paper, we investigate the USRP internal working verilog modules for signal generation. The transmit section is reprogrammed inside the FPGAto free the bandwidth of USB to motherboard of USPR [4], [5], [6], [7]. USRP

M. Singh (✉) · S.K. Bhoi · P.M. Khilar
Department of Computer Science and Engineering, National Institute of Technology,
Rourkela 769008, Odisha, India
e-mail: muneshpal24@gmail.com

S.K. Bhoi
e-mail: souravbhoi@gmail.com

P.M. Khilar
e-mail: pmkhilar@nitrkl.ac.in

P.K. Sa et al. (eds.), *Progress in Intelligent Computing Techniques: Theory, Practice, and Applications*, Advances in Intelligent Systems
and Computing 719, DOI 10.1007/978-981-10-3376-6_60

is
primarily designed to implement the telecommunication application that is built in GNU Radio. It takes the advantage of flexible digital technology of AD9860 DSP chips and the Altera cyclone FPGA. These components perform the high speed sampling, digital down-conversion, digital up-conversion, and capture large variety of radio frequency signal. Our work is motivated with the merits of USRP and its flexible FPGA with telecommunication capability. In this work, we have resolve the issues related to the bandwidth limitation of USB-based USRP for high bandwidth processing. The prototype platform use RFX2400 transceiver that work in full duplex mode with 40 MHz of bandwidth, operate at the frequency of around (2.3–2.9 GHz), and output power of 50 mW [4]. These frequencies are within the operating range of S-band and C-band radars. The prototype radar is built by recompiling the internal verilog module with transmit section logic for FMCW signal generation. The tool used to recompile the internal verilog module is Quartus tool. Using Quartus tool, we can change or design the inside core module logic according to our need. Various useful module already build inside the FPGA written in hardware define language (verilog). The modules performs various function such as CORDIC (Coordinate Rotation Digital Computer), low pass filtering, decimation, and interpolation. The flexibility of FPGA realize the high bandwidth processing on this platform that is useful for short distance FMCW radar. The entire transmit section is built inside the FPGA for signal generation, modulation, and multiplier to generate the beat frequency data. The acquired beat frequency data is transmitted from USB to host machine for further processing. To avoiding the clutter, we use log periodic direction antenna work at frequency range of 850 MHz–6.5 GHz.

The remaining part of this paper is organized as follows. Section 2 presents the background of FMCW radar. Section 3 presents the realization of FMCW radar using USRP platform. Section 4 presents the experimental setup and results, Sect. 5 presents the conclusion.

2 Background of FMCW Radar

2.1 Basic Principle of FMCW

The sawtooth (or ramp) waveform provides only positive frequency sweep, which makes the control and electronic tuning uncomplicated. In FMCW radar, the transmitted RF waveform is linearly swept in frequency. The received signal is then mixed with the transmitted signal to generate the delay caused by time of flight of the reflected signal. Mixing of transmitted signal and reflected signal generate the frequency difference that corresponds to different target component frequency.

$$f_{TXOut} = f_{SF0} + K_I * t,\ 0 \leq t < T, \tag{1}$$

where f_{SF0} is the initializing frequency, T is the frequency sweet time, and K_I is the slope of frequency increment, which is calculated as

$$K_I = \frac{BW}{T}, \tag{2}$$

where BW is the bandwidth of frequency sweep. The round trip time taken from transmitter to target is calculated as

$$\Delta t = 2\frac{d}{c}, \tag{3}$$

where d is the distance from antenna to the target and c is the speed of light. The delay caused by the fight of the transmitted signal from source to the target and back to source again is shown as follows:

$$f_{Received} = f_{SF0} + K_I * (t - \Delta t), \Delta t \le t < T + \Delta t \tag{4}$$

Due to the delay Δt in frequency between $f_{Received}$ and f_{SF0}, the beat frequency Δf is calculated as follows:

$$\Delta f = \frac{BW}{T}.2\frac{d}{c} \tag{5}$$

Different targets are distinguished by different echoes. Each target echo has a unique intermediate frequency (IF) and this frequency component is extracted using Fast Fourier Transform (FFT). The FFT of the sample signal generates different peaks corresponding to the IF of the target standing at particular distances. The accuracy of the target range depends on the sampling rate and sweep rate [5], [6], [7]. According to nyquist, the sampling rate must be twice of the bandwidth to accurately extract the information from the return signal. However, FMCW radar provides the ability to detect the target even without considering the nyquist criterion. There are two ways to define the sampling rate without going beyond the capacity of digital to analog convertor.

- For a complex sampled signal, the sample rate can be set same as the bandwidth.
- Otherwise, sampling rate can be set to the twice of the maximum beat frequency.

The range resolution ΔR of the FMCW radar is given by the equation as follows:

$$\Delta R = \frac{c}{2B} \tag{6}$$

The frequency of beat signal is proportional to the range and its phase containing the Doppler information. In this paper, we only determine the range of the target from beat frequency data.

3 Software Defined Radar

The prototype short-range FMCW radar is divided into three sections such as USB bandwidth limitation, transmit chain section, and receiving chain section.

3.1 USB Bandwidth Limitation

USRP has USB 2.0 interface to host and it can support a maximum speed up-to 32 MB/s [4]. For the complex samples of 16 bit for real part I and 16 bit for imaginary part Q, further divides the maximum bandwidth of USB to 8 MB/s. However, our prototype FMCW radar transmit section is independent from USB link, hence the entire bandwidth of USB is only utilized by the receiving section of the USRP. Now, the maximum bandwidth of the USB is 16 MB/s, which is more than enough to receive the data from USRP to host.

3.2 Transmit Section

The proposed technique uses the digital approach called direct digital synthesis (DDS) to generate the chrip signal. DDS generates highly accurate harmonically pure digital representation of the signal by accumulating phase change at much higher frequency [4]. A digital phase accumulator increments a constant phase in each cycle of the reference clock. In this system, the output frequency is the function of clock frequency f_{clk}, the length (in bits) of phase accumulator N and the phase increment $\Delta\phi$. The output of the phase accumulator is sawtooth waveform that signifies the linearly changing phase of a sinusoidal signal.

The CORDIC architecture is defined inside the FPGA to perform the elementary rotations, which only requires shift and add operations to generate the sine and cosine waveform [4]. CORDIC generated waveform is directly fed into the digital to analog convertor (DAC). Figure 1a shows the register transfer level (RTL) view of the radar logic for signal generation and modulation, which is complied using the Quartus tool. The transmit section logic is verified using GTKWAVE, as shown in Fig. 1b.

3.3 Receiver Section

The receiving chain of the USRP, down convert the reflected signal and multiplied with the transmitted signal to generate the beat frequency data. The complex multiplier module is build inside the FPGA to multiply the transmitted signal with received echo. From the receiving echo, target position is calculated by correlating the beat frequency corresponding to the distances.

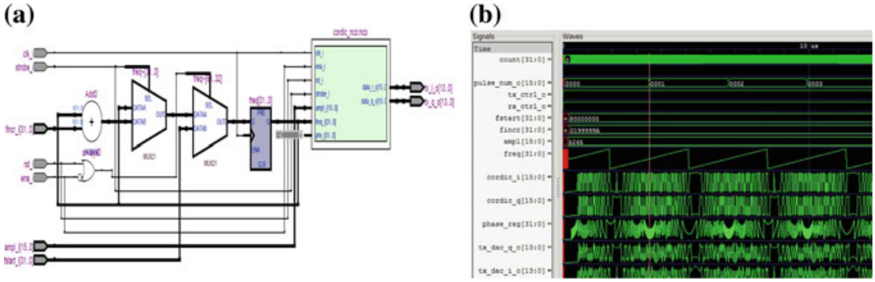


Fig. 1 Radar build logic: **a** designed RTL view of transmit section built inside the FPGA. **b** Logic verified using GTKWAVE

4 Experimental Setup and Results

The FMCW radar prototype specifications are shown in Table 1. To test the applicability of USRP platform as a short range FMCW radar, an experimental setup is performed. Test experiments are carried out on terrace environment of area 100×50 m^2 approximately. The target is standing beyond the range resolution of the radar. To avoid the clutter of the surrounding environment, the log periodic directional antenna is used. In Fig. 2, we have shown the prototype experimental platform for testing the functionality of FMCW radar build inside the USRP. In our experiment, we stand a target at a distance of 20 m and estimate the beat frequency at this distance. To avoid the processing delay of internal circuitry of USRP, we skip the 1000 samples from the received beat frequency data. Later, the time domain data of beat frequency shown in Fig. 3a is transformed into the frequency domain using FFT, as shown in Fig. 3b. From Fig. 3b, it is observed that a peak is found at a same distance of 20 m, which is corresponding to the beat frequency of distance 20 m where target is placed.

Table 1 System design parameters

Parameters	Values
Center frequency	2.4 GHz
Bandwidth	15 MHz
Sampling rate	15 MHz
Sweep period	5 μs
Sweep bandwidth	15 MHz
Range resolution	10 m
Maximum range	1.5 km

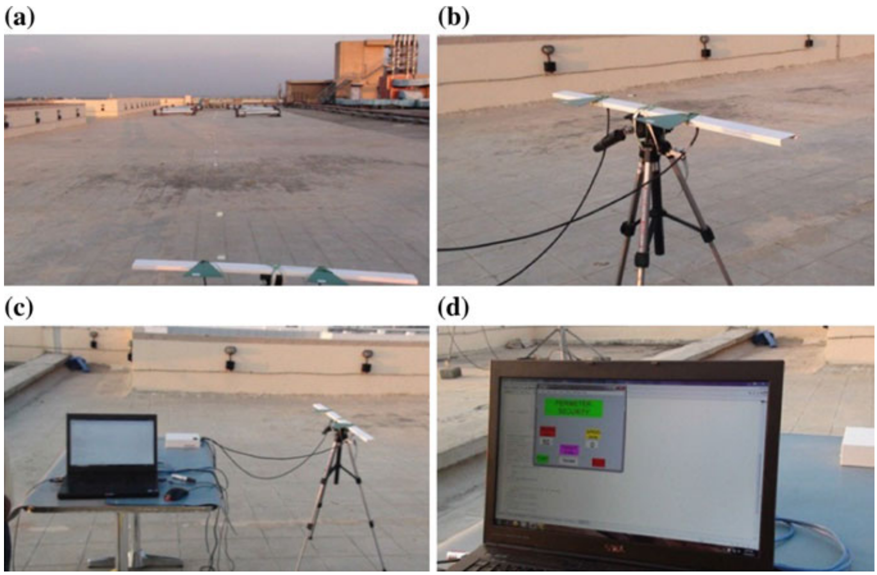


Fig. 2 Experimental platform: **a** testing range of 100 m approximately. **b** Log periodic directional antenna setup. **c** Entire setup for experimental analysis. **d** GUI for displaying the target distance

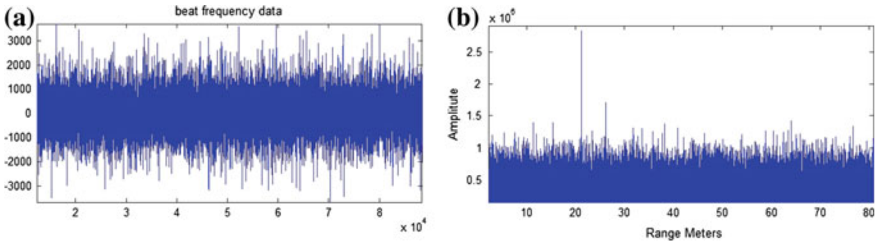


Fig. 3 Experimentally acquired data: **a** Beat frequency data in time domain. **b** Beat frequency data in frequency domain

5 Conclusion

Software defined radar gives a high level of programmability and functionality rather than the classical radar, therefore appearing as a ideal and low cost solution to construct an effective ATR system. In this particular experiment, we demonstrate the applicability of this platform for high bandwidth processing. The result shows the ability of USRP platform to design the short distance FMCW radar.

References

1. Skolnik, M. Introduction to Radar Systems.New York: McGraw-Hill, 2000.
2. Singh, Munesh, and Pabitra Mohan Khilar. (2016). Mobile beacon based range free localization method for wireless sensor networks. Wireless Networks: pp. 1–16. Springer.
3. Singh, M. and Khilar, P.M. (2015), An analytical geometric range free localization scheme based on mobile beacon points in wireless sensor network. Wireless Networks, pp. 1–14. Springer.
4. Hamza, F. A. (2008). The USRP under 1.5 x magnifying lens!. https://microembedded.googlecode.com/files/USRP_Documentation.pdf.
5. http://siversima.com/wp-content/uploads/FMCW-Radar-App-Note.pdf , Accessed 4th July 2015.
6. Pancik, J., and Pancik, M. (2015, April). Hardware and software front-end based on the USRP for experimental X-band Synthetic Aperture Radar. In Radioelektronika (RADIOELEKTRONIKA), 2015 25th International Conference (pp. 154–159). IEEE.
7. Prasetiadi, A. E., and and Suksmono, A. B. (2012, April). A simple delay compensation system in software-defined Frequency Modulated Continuous (FMCW) Radar. In European Wireless, 2012. EW. 18th European Wireless Conference (pp. 1–4). VDE.

Erratum to: A Browser-Based Distributed Framework for Content Sharing and Student Collaboration

Shikhar Vashishth, Yash Sinha and K. Haribabu

Erratum to:
Chapter 12 in: P.K. Sa et al. (eds.), *Progress in Intelligent Computing Techniques: Theory, Practice, and Applications*, Advances in Intelligent Systems and Computing 719, https://doi.org/10.1007/978-981-10-3376-6_12

In the original version of the book, the author name "K. Hari Babu" has to be changed to "K. Haribabu" in Chapter 12, which is a belated correction. The erratum chapter and the book have been updated with the change.

The updated online version of this chapter can be found at
https://doi.org/10.1007/978-981-10-3376-6_12

P.K. Sa et al. (eds.), *Progress in Intelligent Computing Techniques: Theory, Practice, and Applications*, Advances in Intelligent Systems and Computing 719, https://doi.org/10.1007/978-981-10-3376-6_61

Erratum to: Progress in Intelligent Computing Techniques: Theory, Practice, and Applications

Pankaj Kumar Sa, Manmath Narayan Sahoo, M. Murugappan, Yulei Wu and Banshidhar Majhi

Erratum to:
P.K. Sa et al. (eds.), *Progress in Intelligent Computing Techniques: Theory, Practice, and Applications*, Advances in Intelligent Systems and Computing 719, https://doi.org/10.1007/978-981-10-3376-6

The original version of the book was inadvertently published with incorrect volume number 519 which has been now corrected as 719.

The updated online version of this book can be found at
https://doi.org/10.1007/978-981-10-3376-6

P.K. Sa et al. (eds.), *Progress in Intelligent Computing Techniques: Theory, Practice, and Applications*, Advances in Intelligent Systems and Computing 719, https://doi.org/10.1007/978-981-10-3376-6_62

Author Index

P.K. Sa et al. (eds.), *Progress in Intelligent Computing Techniques: Theory, Practice, and Applications*, Advances in Intelligent Systems and Computing 719, DOI 10.1007/978-981-10-3376-6