Computer Science & Information Technology 26

Sundarapandian Vaidyanathan
Sangeeta Dhananjay Jadhav (Eds)

Computer Science & Information Technology

Fourth International Conference on Computational Science, Engineering and Information Technology (CCSEIT 2014)
Pune, India, August 08 ~ 09 - 2014

AIRCC

Volume Editors

Sundarapandian Vaidyanathan,
R & D Centre,
Vel Tech University, India
E-mail: sundarvtu@gmail.com

Sangeeta Dhananjay Jadhav
Asian Institute of technology, India
E-mail: djsangeeta@rediffmail.com

ISSN: 2231 - 5403
ISBN: 978-1-921987-10-6
DOI : 10.5121/csit.2014.4801 - 10.5121/csit.2014.4809

Typesetting: Camera-ready by author, data conversion by NnN Net Solutions Private Ltd., Chennai, India

Preface

Fourth International Conference on Computational Science, Engineering and Information Technology (CCSEIT 2014) was held in Pune, India, during August 08~09, 2014. Third International Conference on Mobile & Wireless Networks (MoWiN 2014), First International Conference on Artificial Intelligence and Applications (AIAP 2014), First International Conference on Bioinformatics and Biosciences (ICBB 2014) and First International Conference on Data Mining and Database (DMDB 2014) were collocated with the CCSEIT-2014. The conferences attracted many local and international delegates, presenting a balanced mixture of intellect from the East and from the West.

The goal of this conference series is to bring together researchers and practitioners from academia and industry to focus on understanding computer science and information technology and to establish new collaborations in these areas. Authors are invited to contribute to the conference by submitting articles that illustrate research results, projects, survey work and industrial experiences describing significant advances in all areas of computer science and information technology.

The CCSEIT-2014, MoWiN-2014, AIAP-2014, ICBB-2014, DMDB-2014 Committees rigorously invited submissions for many months from researchers, scientists, engineers, students and practitioners related to the relevant themes and tracks of the workshop. This effort guaranteed submissions from an unparalleled number of internationally recognized top-level researchers. All the submissions underwent a strenuous peer review process which comprised expert reviewers. These reviewers were selected from a talented pool of Technical Committee members and external reviewers on the basis of their expertise. The papers were then reviewed based on their contributions, technical content, originality and clarity. The entire process, which includes the submission, review and acceptance processes, was done electronically. All these efforts undertaken by the Organizing and Technical Committees led to an exciting, rich and a high quality technical conference program, which featured high-impact presentations for all attendees to enjoy, appreciate and expand their expertise in the latest developments in computer network and communications research.

In closing, CCSEIT-2014, MoWiN-2014, AIAP-2014, ICBB-2014, DMDB-2014 brought together researchers, scientists, engineers, students and practitioners to exchange and share their experiences, new ideas and research results in all aspects of the main workshop themes and tracks, and to discuss the practical challenges encountered and the solutions adopted. The book is organized as a collection of papers from the CCSEIT-2014, MoWiN-2014, AIAP-2014, ICBB-2014, DMDB-2014.

We would like to thank the General and Program Chairs, organization staff, the members of the Technical Program Committees and external reviewers for their excellent and tireless work. We sincerely wish that all attendees benefited scientifically from the conference and wish them every success in their research. It is the humble wish of the conference organizers that the professional dialogue among the researchers, scientists, engineers, students and educators continues beyond the event and that the friendships and collaborations forged will linger and prosper for many years to come.

Sundarapandian Vaidyanathan
Sangeeta Dhananjay Jadhav

Organization

Program Committee Members

Abd El-Aziz Ahmed	Anna's University, Egypt
Abdolreza Hatamlou	Islamic Azad University, Iran
Abdulmohsen Algarni	King Khalid University, Saudi Arabia
Abdurrahman Celebi	Beder University, Albania
Aiden B. Lee	Qualcom Inc, USA
Ali Abid D. Al-Zuky	Mustansiriyah University, Iraq
Apai	Universiti Malaysia Perlis, Malaysia
Batuhan AYHAN	Marmara University, Turkey
Ben Khayut	Intelligence Decisions Technologies, Israel
BenZidane Moh	University of Constantine, Algeria
Chin-Chih Chang	Chung Hua University, Taiwan
Dac-Nhuong Le	Haiphong University, Vietnam
Dammak Nouha	MIRACL laboratory, Tunisia
Denivaldo Lopes	Federal University of Maranhão, Brazil
Deperlioglu Omer	Afyon Kocatepe University, Turkey
Dires Fasil Fenta	University of Gondar, Ethiopia
Dongchen Li	Peking University, China
Emilio UR	University of La Rioja, Spain
Farhad Soleimanian	Hacettepe University, Turkey
Frank Moisiadis	University of Notre Dame Australia, Australia
Gaurav Ojha	Indian Institute of Information Technology and Management, India.
Girija. Chetty	Univeraity of Canberra, Australia
Hacene Belhadef	University of Constantine, Algeria
Hamadouche M	University of Blida, Algeria
Hamdi Hassen	Taibah University, Saudi Arabia
Hazem Alnajjar	Misurata university, Libya
Hyung-Woo Lee	Hanshin University, Korea
John Tengviel	Sunyani Polytechnic, Ghana
Keneilwe Zuva	University of Botswana, Botswana
Khaled Merit	University of mascara, Algeria
LI Zhongqi	Qualcomm Inc, USA
Liyakath Unisa	Prince Sultan University, Saudi Arabia
Mahesh Manik Kumbhar	FTC, COER, Sangola
Masoud Ziabari	Mehr Aeen University, Iran
Md Shohidul Islam	University of Ulsan, South Korea
Mohamed Ali Mahjoub	National Engineering School of Sousse, Tunisia
Mohamed el boukhari	University Mohamed First, Morocco
Moses Ekpenyong	University of Uyo, Nigeria
Nasser Tairan	King Khalid University, Saudi Arabia
Natarajan Meghanathan	Jackson State University, USA
Neda Darvish	Islamic Azad University, Iran
Neela Madheswari A	KMEA Engineering College, India

Nguyen Dinh, Thuc	University of Science, VNU-HCMC, Vietnam
Nishant Doshi	SVNIT, India
Pradnya Kulkarni	Federation University, Australia
Rafah M. Almuttairi	University of Babylon, Iraq
Rao	Hewlett-Packard, USA
Reza Ebrahimi Atani	University of Guilan, Iran
Saad M. Darwish	Alexandria University, Egypt
Sean McGerty	University of Notre Dame Australia, Australia
Seyyed Mohammadreza Farshchi	Tehran University, Iran
Sid Kulkarni	Federation University, Australia
Sundarapandian Vaidyanathan	Vel Tech University, India
T.V.Prasad	Visvodaya Technical Academy, India
Tad Gonsalves	Sophia University, Japan
Tinatin Mshvidobadze	Gori University, Georgia
Utku Kose	Usak University, Turkey
V.S.Dharun	Noorul Islam University, India
William R Simpson	Institute for Defense Analyses, USA
Xiantao Zhang	Peking University, China

Technically Sponsored by

Computer Science & Information Technology Community (CSITC)

Networks & Communications Community (NCC)

Digital Signal & Image Processing Community (DSIPC)

Organized By

Academy & Industry Research Collaboration Center (AIRCC)

TABLE OF CONTENTS

Fourth International Conference on Computational Science, Engineering and Information Technology (CCSEIT 2014)

First International Conference on Data Mining and Database (DMDB 2014)

First International Conference on Bioinformatics and Biosciences (ICBB 2014)

Third International Conference on Mobile & Wireless Networks (MoWiN 2014)

First International Conference on Artificial Intelligence and Applications (AIAP 2014)

SECURITY WEAKNESSES ON A MUTUAL AUTHENTICATION AND KEY AGREEMENT SCHEME IN GLOBAL MOBILE NETWORKS

Prosanta Gope[1] and Tzonelih Hwang[2]

[1]National Cheng Kung University, Tainan, Taiwan, R.O.C
prosanta.nitdgp@gmail.com
[2]National Cheng Kung University, Tainan, Taiwan, R.O.C
hwangtl@ismail.csie.ncku.edu.tw

ABSTRACT

User mobility is a feature that raises many issues related to security. One of them is the disclosure of a mobile user's real identity during the authentication process, or the other procedures specific to global mobile networks (GLOMONET). Such disclosure allows an unauthorized third-party to track the mobile user's movements and current whereabouts. In this article, we address some problems of mutual authentication and key agreement with user anonymity for GLOMONET. Recently, Qi et al. proposed such scheme, which is claimed to be a slight modification of He et al.'s protocol based on smart card. However, we reveal that both the schemes still suffer from certain weaknesses which have been overlooked previously and thus they cannot achieve desired security.

KEYWORDS

Authentication, Anonymity, Roaming, Privacy, Untraceability, Smart card, Global mobile network.

1. INTRODUCTION

Global mobile network (GLOMONET) is a useful networking environment which permits a mobile user to access the services provided by the home agent (HA) in a foreign network (FA). For securing the communication conducted over GLOMONETs, it is important to provide a way for authenticating mobile users in an anonymous manner. Besides, in the design of an efficient authentication scheme for roaming services in GLOMONET, mutual authentication must be supported to prevent any illegal use of resources and to ensure that mobile users are connected to the trusted network. In order to do so, the authentication scheme should have ability to resist various kinds of attacks or any forgery attempts. For accomplishing these goals, many authentication and key agreement schemes have been proposed with anonymity for roaming services in global mobile networks [1-7]. Particularly, in 2004, Zhu et al. proposed a wireless security protocol based on smart card and featuring user anonymity [1]. Unfortunately, Lee and Hwang [2] pointed out in 2006 that Zhu and Ma's protocol's [1] does not achieve mutual authentication and is also subjected to the forgery attack.

Sundarapandian et al. (Eds) : CCSEIT, DMDB, ICBB, MoWiN, AIAP - 2014
pp. 01–07, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4801

Lee et al. also proposed a slightly modified version of Zhu et al's protocol so as to remedy the identified shortcomings. However, in [3], it was shown that the Zhu et al.'s scheme and Lee and et al.'s scheme fails to provide user anonymity, and Wu, Lee and Tsaur proposed an enhanced scheme by providing an effective remedy. Independently, in [4], Chang et al. showed that Lee et al.'s scheme cannot provide user anonymity under the forgery attack and also proposed an enhanced authentication scheme. Unfortunately, Youn et al. found that the scheme of [4] fails to achieve user anonymity under four attack strategies [5]. Thereafter, He et al. proposed an improved scheme [6] based on the concept of pseudonym. However, the scheme is considered to be economically impractical because of the extraction of parameters from the private space of the smart card. Besides, recently, Qi et al. [7] pointed out some other drawbacks of the He et al. scheme and they proposed an improved authentication protocol for GLOMONET environment. However, in this article, we show that both the schemes [6-7] have some serious weaknesses which have been overlooked.

The remainder of this article is organized as follows. Section 2 reviews the protocol of [7] and whose weaknesses are pinpointed in Section 3. Finally, a concluding remark is given in Section 4. The abbreviations and cryptographic functions used in this article are defined in Table 1.

Table 1. Notation and Abbreviations.

Notation	Description
MS	Mobile station/User
FA	Foreign agent
HA	Home agent
ID_M	Identity of a mobile user
ID_f	Identity of a foreign agent
ID_h	Identity of a home agent
PSW_M	Password of the mobile user
$h(.)$	One-way hash function
$\oplus$	Exclusive-OR operation
P	Concatenation operation

2. REVIEW OF QI ET AL.'S SCHEME

Qi et al. scheme [7], which is claimed to be a slight modification of, but a security enhancement on He et al.'s scheme [6], consists of three phases. In Phase I, the home agent (HA) security issues a smart card to a mobile user MS. In Phase II, mutual authentication between MS and a foreign agent (FA) is performed under the supervision of the home agent (HA). After the successful authentication, a legitimate MS can access the wireless service from the FA, and establish a session key between them. In Phase III, MS can renew his/her password. It is assumed that each foreign agent FA shares a long-term secret key K_{fh} with home agent HA.

2.1 Assumption on Quadratic Residue

Conceive, p and q two large primes, from that we calculate $n = p*q$. Now, if $y = x^2 \bmod n$ has a solution, in other words, there exists a square root for y, then y is a quadratic residue of mod n. Therefore, we can represent the set of all quadratic residue numbers in [1, n-1] by QR_n. Based on

the quadratic residue assumption, which states that for any $y \in QR_n$, it is difficult to figure out the value of x without having any prior knowledge of p and q because of the difficulty of factoring n[8] into two prime factors which is indeed a difficult task.

2.2 Phase I: Registration Phase

When a mobile user wants to register at the home agent HA, the user has to submit a request to the home agent, and then home agent will issue a smart card with the related information to the user. In this regard, MS at first submits his/her claimed identity ID_M to HA in via a secure channel. After receiving the request from MS, the home agent HA generates a secret random number x and computes $K_{ms} = h(ID_M \| x)$ and then the system (HA) generates two large primes p and q and computes n = p*q. Finally, HA personalized a smart card with $h(.)$, K_{ms}, and n, and issues it to MS via a secure channel and then stores ID_M, and K_{ms} for further communication. Hereafter, MS computes $K_{ms}^{*} = h(ID_M \| x) \oplus h(ID_M \| PSW_M)$ and replaces K_{ms}^{*} with and holds $h(.)$, K_{ms}^{*}, and n for further communication.

2.3 Phase II: Login and Authentication Phase

Once enrolled by HA, when MS visits a foreign network managed by the FA, he/she needs to authenticate himself/herself to FA in order to show that he/she is a legitimate subscriber of his/her home network managed by HA. In this phase FA authenticates MS under the assistance of HA, and issues a session key SK. The steps of this phase are outlined in Fig. 1 and explained as follows.

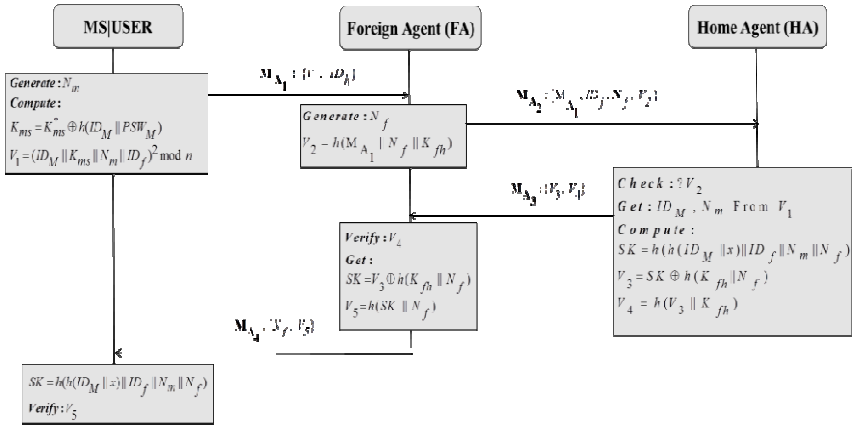


Figure.1 Login and Authentication Phase of Qi et al.'s Scheme

Step 1 $M_{A_1} : MS \rightarrow FA: \{V_1, ID_h\}$.

MS submits his/her identity ID_M and password PSW_M to the smart card and computes $K_{ms} = K_{ms}^{*} \oplus h(ID_M \| PSW_M)$. Hereafter, MS generates a random number N_m and derives $V_1 = (ID_M \| K_{ms} \| N_m \| ID_f)^2 \bmod n$. Finally, MS sends the login message M_{A_1} to FA.

Step 2 $M_{A_2} : FA \rightarrow HA: \{M_{A_1}, ID_f, N_f, V_2\}$.

After receiving the login message M_{A_1}, FA generates a random number N_f and computes $V_2 = h(M_{A_1} \| N_f \| K_{fh})$. Hereafter, it requests the HA by sending it's claimed identity ID_f, the nonce N_f, V_2, in addition to those of MS.

Step 3 M_{A_3} : HA $\rightarrow$ FA : $\{V_3, V_4\}$.

Upon receiving the request from FA, the home agent HA checks ID_f and then computes and verifies whether V_2 is equal to $h(M_{A_1} \| N_f \| K_{fh})$ or not. After successful verification, HA solves V_1 by using the Chinese Remainder Theorem [8-10] with p and q and get ID_M, K_{ms}, N_m, and ID_f. Thereafter, HA computes $h(ID_M \| x)$ and verifies with the received K_{ms}. If the verification is successful then HA considers the MS as a legitimate mobile subscriber. Hereafter, HA computes $SK = h(h(ID_M \| x) \| ID_f \| N_m \| N_f)$, , $V_3 = SK \oplus h(K_{fh} \| N_f)$, $V_4 = h(V_3 \| K_{fh})$ where SK denotes the session key between MS and FA. Finally, HA forms M_{A_3} and sends it to FA.

Step 4 M_{A_4} : FA $\rightarrow$ MS : $\{N_f, V_5\}$.

Upon receiving M_{A_3}, FA checks whether V_4 is equal to $h(V_3 \| K_{fh})$ or not. After successful verification, FA computes $SK = V_3 \oplus h(K_{fh} \| N_f)$, $V_5 = h(SK \| N_f)$ and sends M_{A_4} to MS. After receiving M_{A_4} from FA, MS at first computes $SK = V_3 \oplus h(K_{fh} \| N_f)$ and then verifies V_5 is equal to $h(SK \| N_f)$ or not. If it is true, then MS establishes a SK with FA; otherwise authentication fails.

2.4 Phase III: Password Renewal Phase

In this scheme, a mobile user can freely change his/her password on the smart card without the help of the home agent HA. Now, when mobile user MS with a smart card wants to change the password of the smart card, MS makes a request to the smart card, and then inputs the old password PSW_M and the new password PSW_M to the smart card. Then the smart card recovers $K_{ms} = K_{ms}^{*} \oplus h(ID_M \| PSW_M)$ and derives $K_{ms}^{**} = K_{ms}^{*} \oplus h(ID_M \| PSW_M^{*})$ .Finally, stores the K_{ms}^{**} in place of K_{ms}^{*}.

3. SECURITY WEAKNESSES IN QI ET AL.'S PROTOCOL

During the cryptanalysis of the He et al.'s scheme, Qi et al. shown that the He et al's scheme is highly insecure because of the several attacks like "mobile tracking and identity guessing attack", "offline guessing attack" etc. In this regard, the adversary needs to perform some exhaustive guess operations and through which he/she needs to figure out one unknown parameter from a relation, where other parameters in the relation are publicly known (unencrypted data). Unfortunately, while designing their improved scheme [7], like He et al., Qi et al. also overlooked that issue in some cases. As a consequence of that Qi et al.'s improved scheme still has several serious deficiencies (shown below).

3.1 Revealing of long-term Secret Key and Session Key

Consider an adversary $\mathcal{A}$ has control over the communicating messages transmitted over open networks. Precisely, the adversary $\mathcal{A}$ has the capability to intercept the messages flowing through the mobile network. Now, in the login and authentication phase of the Qi et al.'s scheme, after the successful verification of the mobile user, as well as the foreign network when, the home agent HA sends the response message M_{A_3} to FA. We assume that the adversary has intercepted that message. Therefore, $\mathcal{A}$ receives both V_3, V_4, where $V_4 = h(V_3 \| K_{fh})$. In this relation, only K_{fh} is unknown to the adversary $\mathcal{A}$. Therefore, by executing an exhaustive search operation, he/she can easily figure out the long-term shared secret key K_{fh}, which is indeed a serious concern. As, this will not only affect that particular mobile subscriber, at the same time it also compromises the security of other mobile users who received their smart card from that particular home agent and willing to roam over through the area covered by that particular foreign agent whose secret key K_{fh} has been compromised. In this case, after acquiring that secret key, the adversary can perform any kind of forgery attempt and even can share this secret key with a dishonest foreign agent who can exploit it with its superior capabilities and that may even annoy the mobile subscriber with billing problem. Now, we consider that the adversary $\mathcal{A}$ eavesdrops the communication between MS and FA, in other words, the adversary has intercepted the response message M_{A_4}, which is sent by the foreign agent FA to MS. In this regard, using the parameters N_f, V_5, where $V_5 = h(SK \| N_f)$ and executing the exhaustive search operation, the adversary can easily get the session key *SK* every time, which is also a serious apprehension. The similar problems can also be profound in the login and authentication phase of the He et al.'s scheme.

.

3.2 Vulnerable to Known Session Key Attack

It is obvious that known session key attack is a serious threat against any session key establishing schemes. A protocol is called secure against known session key attacks if a revealed session key does not influence on the security of other session keys. In other words, if past session keys are compromised, it should not allow an adversary to compromise future session key or any even any other session keys earlier than that one. In this way, a protocol can also compromise its backward and forward secrecy. Where, by backward secrecy, we mean that a compromise of any session key should not compromise any earlier key. While forward secrecy implies that a compromise of the current session key should not compromise any future key. However, unfortunately, the Qi et al.'s scheme cannot ensure the security against known session key attack. If a session key established between MS and FA is revealed to an adversary, then the adversary may target the relation $SK = V_3 \oplus h(K_{fh} \| N_f)$, where the parameters V_3 and N_f are public. Therefore, by executing the exhaustive search operation, the adversary $\mathcal{A}$ can figure out K_{fh}. Now, by using that the adversary can easily acquire all the past and future session keys. Accordingly, Qi et al.'s scheme cannot insure any forward and backward secrecy. Moreover, this problem also persists in the login and authentication phase of the He et al.'s scheme.

3.3 Unsuccessful Key-Agreement

In the login and authentication phase of [7], after acquiring the secret key K_{fh} by executing an exhaustive search operation presented in Sect. 3.1 and 3.2, if the adversary $\mathcal{A}$ eavesdrops the

message M_{A_3} and then attempts to modify V_3 to $V_3^{'}$, and using that, alters V_4 to $V_4^{'}$ and then sends $\{V_3^{'}, V_4^{'}\}$ to the foreign agent. However, unfortunately, the foreign agent cannot comprehend that alternation. Because, when the foreign agent verifies the relation $h(V_3^{'} \| K_{fh})$ with $V_4^{'}$, it seems to be perfect for the system (FA). In this way, an attacker can resist a legitimate foreign agent FA to produce the valid session key and eventually that makes the key-agreement unsuccessful. In fact, the similar problem can also be profound in [6].

3.4 Logical Error during Renewal of Password

Apart from the aforesaid serious security issues, the improved scheme proposed by Qi et al. also encompasses one logical mistake during the execution of the password renewal phase. After retrieving the original K_{ms} through $K_{ms} = K_{ms}^{*} \oplus h(ID_M \| PSW_M)$, the smart card updates the K_{ms}^{**} with the relation $K_{ms}^{**} = K_{ms}^{*} \oplus h(ID_M \| PSW_M^{*})$. Where K_{ms}^{*} denotes the previously updated key. In this way, MS forgets the original K_{ms} which was given by the home agent HA during registration. . As a consequence of that, after the execution of the password renewal phase, if MS wants to acquire roaming services in that case during authentication the home agent cannot match the K_{ms} with the relation $h(ID_M \| x)$. This will surely compel both the MS and HA to execute the registration phase once again, which is not desired at all.

4. CONCLUSION

Recently, Qi et al. proposed a authentication and key agreement protocol featuring user anonymity. In this article, we have demonstrated certain deficiencies found in Qi et al.'s proposed authentication scheme. Where, we have found that the scheme presented by Qi. et al. is vulnerable to known session key attack, forgery attacks etc. Nevertheless, the proposed scheme has also committed some logical errors during the password renewal phase.

ACKNOWLEDGEMENTS

This work is financially supported by the National Science Council of Republic of China (Taiwan), under Contract No. NSC MOST 103-2221-E-006 -177 -. The authors would like to thank the National Science Council of Republic of China for their benign supports

REFERENCES

[1] Zhu, J. & Ma, J. (2004). A new authentication scheme with anonymity for wireless environments, IEEE Transactions on Consumer Electronics, 50(1) 230-234.

[2] Lee,C., Hwang, M. S., & Liao, I. E. (2006). Security enhancement on a new authentication scheme with anonymity for wireless environment, IEEE Transactions on Industrial Electronics, 53(5), 1683-1687.

[3] Wu, C. Lee, W. B. & Tsaur, W. J. (2008). A secure authentication scheme with anonymity for wireless communications, IEEE Communication Letters, 12(10), 722-723.

[4] Cheng, C.C. Lee, C.Y. & Chiu, Y.C. (2009) Enhance authentication scheme with anonymity for roaming service in global mobility networks, Computer Communications, 32, 611-618.

[5] Youn, T. Y., Park, T. H., & Lim. (2009). Weaknesses in an anonymous authentication scheme for roaming service in global mobile networks, IEEE Communication Letters, 13(7), 471-473.

[6] He, D., Ma, M., Chen, C., and Bu J. (2011). Design and validation of an efficient authentication scheme with anonymity for roaming service in global mobility networks, Wireless Personal Communications, 61, 465-476.
[7] Jiang Q., Ma, J., Li, G., Yang, L. (2013), An enhanced authentication scheme with privacy preservation for roaming services in global mobility networks, Wireless Personal Communications, 68, 1477-1491.
[8] Rosen, K. (1988). Elementary number theory and its applications, Addision Wesley.
[9] Stalling, W. (2000). Cryptography and network security principles and practice, Prentice Hall.
[10] Trappe, W., and Washington, C., L. (2006). Introduction to cryptography with coding theory, Prentice Hall.
[11] Chang, C., Lee, J. & Chang, Y. (2005). Efficient authentication protocols of GSM. Computer Communications, 28 (8), 921-928.
[12] C.C Lo, Y.J.Chen.(1997). Secure communication mechanisms for GSM networks. In Proceedings of the IEEE transactions on Consumer Electronics 45, 1074-1080.
[13] T-F Lee, C. C Chang and T. Hwang, Private Authentication Techniques for the Global Mobility Network, ' Wireless Personal Communications, Vol 35, No 4, Jan 2005, pp. 329-336
[14] Hwang, T. Gope, P. Provably secure mutual authentication and key exchange scheme for expeditious mobile communication through synchronously one-time Secrets, Wireless Personal Communications, DOI. 10.1007/s11277-013-1501-5, 2013
[15] TS 33.102: Security architecture, version 4.2.0, released 4. Third Generation Partnership Project-Technical Specification Group, 2001.
[16] TR 33.902: Formal analysis of the 3G authentication protocol. Third Generation Partnership Project-Authentication and Key Agreement (AKA), 2000.

AUTHOR BIOGRAPHIES

Prosanta Gope received his M.Tech degree in Computer Science and Engineering from National Institute of Technology (NIT), Durgapur, India, in 2009. Currently he has been pursuing his PhD degree in Computer Science and Information Engineering, National Cheng Kung University (NCKU), Tainan, Taiwan. His research interests include authentication, authenticated encryption, security in mobile communication and cloud computing.

Tzonelih Hwang received the M.S. and Ph.D. degrees in Computer Science from the University of Southwestern Louisiana, USA, in 1988. He is currently a Distinguished Professor in the department of Computer Science and Information Engineering, National Cheng Kung University (NCKU), Tainan, Taiwan. Dr. Hwang has actively participated in several research activities including as a research scientist at the Center for Advanced Computer Studies, University of Southwestern Louisiana, USA. He is also associated as a vigorous member of the editorial board of some reputable international journals. He has published more than 250 technical papers and holds four patents. His research interests include network and information security, access control systems, error control codes, security in mobile communication and quantum cryptography.

INTENTIONAL BLANK

EYE CONTROLLED SWITCHING USING CIRCULAR HOUGH TRANSFORM

Sagar Lakhmani

Software Engineering Associate,
Accenture Services Private Limited, Pune, India
Sagar.lakhmani@accenture.com

ABSTRACT

The paper presents hands free interface between electrical appliances or devices. This technology is intended to replace conventional switching devices for the use of disabled. It is a new way to interact with the electrical or electronic devices that we use in our daily life. The paper illustrates how the movement of eye cornea and blinking can be used for switching the devices. The basic Circle Detection algorithm is used to determine the position of eye. Eye blinking is used as triggering action which produces binary output through the parallel port of computer. This output is used for switching the particular device.

KEYWORDS

Eye controlled switching, Circular Hough Transform, Circle detection, Hands-free switching

1. INTRODUCTION

The existing appliances used in our daily life such as Fans, Lights, Televisions, and other devices have been used to achieve daily comfort. These electrical appliances cannot be operated by the handicapped persons. In this paper, a computer based input device by human eyes is used for switching of daily use electrical appliances.

The existing Switching methods for interacting with electrical devices are as follows:

1. Voice Based Method, which use user's voice as source input. Voice analysis is used to analyse user's voice and convert into electrical digital signals of ON & OFF. The weakness of this system is presence of noise. Other voices coming from surrounding may affect the system.

2. Hardware switch, the most common method of interacting with electrical devices.

There have been methods involved for controlling of appliances such as Electric Wheelchair, Computer Mouse, etc using Eye tracking. The existing methods using eye movements as input are as follows:

Sundarapandian et al. (Eds) : CCSEIT, DMDB, ICBB, MoWiN, AIAP - 2014
pp. 09–16, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4802

1. Using Infrared Camera, in which the movement of the eye is monitored by high end infrared camera and high precision monitor mounted telescopic cameras. The camera detects the movement of the eye after the calibration process is performed successfully with few training sessions which records reference points given by the system. Once the training is completed, the system uses stored eye positions recorded during the training and current image of the eyes captured by live cameras to plot the expected position using linear interpolation methods. This method required efficient jitter reduction algorithm since the random movement of the user is quite common in practical cases.

2. Using Starburst technique which needs intensive reinforced learning for making the system practically reliable and the cost of implementation is too high. The cameras employed to read the eye movements are highly sophisticated and involve huge investment thereby increasing the cost when production is adopted at a commercial level.

3. Use of KNN classifier to determine various illumination conditions, which is more feasible than lighting compensation processing in real-time implementation. But this is very much dependant on the lighting conditions provided under practical environments and depended on the entire face image acquisition and extracting required features from the frame containing the image of whole face. The extraction of the whole face as the problem domain makes it more complex since the image processing should involve segmentation process to separately isolate the properties of the user's eye.

Many popular eye based appliance handling systems have been developed that generates no feedback when person looks at it. It can be quite distracting to a person when he or she is aware of the gazes and consciously tries to put the efforts to control switching the devices.

2. PROPOSED METHOD

The methodology used is based on the output of the implemented circular Hough Transform for circle detection to sense the image of the eyeball using a camera mounted on a headgear and continuously work upon the video frames to detect the position of the cornea. The position of the cornea determines which device is to be switched ON or OFF. The experimental setup has been constructed to contain a head supported structure containing a wireless webcam which captures the close up live video of the eye. An Object Analysis Toolbox available in latest versions of MATLAB enables experimenters to measure the edges and objects in an image. The imfindcircles function of this toolbox is used to detect the position of the cornea.

The following figure [Figure 1] shows the photograph of the hardware device we have used for the experimental purpose. The device consists of a flexible structure which can be worn as a head gear by the user. The tip of the structure is mounted with a simple webcam which captures the image of the eye.

An IR LED is used to illuminate the eyes and this method is technically known as dark-pupil illumination. The main reason for using this technique is the complete absorption of the IR rays into the cornea of the eyes so as to eliminate specular reflection caused when visible light is used instead. This yields results with excellent accuracy by the image processing module we use to detect the centre of the cornea to obtain its coordinates. The inlet shows the captured image on the computer screen. For better illumination, we have adjusted the position of the camera arm such

that we have minimum blockage of the ambient light present in the laboratory and the shadow of the structure does not fall directly on the cornea.

Figure 1: The hardware device used for capturing the live video of the eye.

The calculation is facilitated by assuming an imaginary square grid placed over the input frames and detect the cell in the grid in which the centre of the eye is currently detected by Circular Hough Transform circle detection algorithm. The centre of the cornea can be calculated by monitoring the co-ordinate of the cell. The following captured image illustrates this.

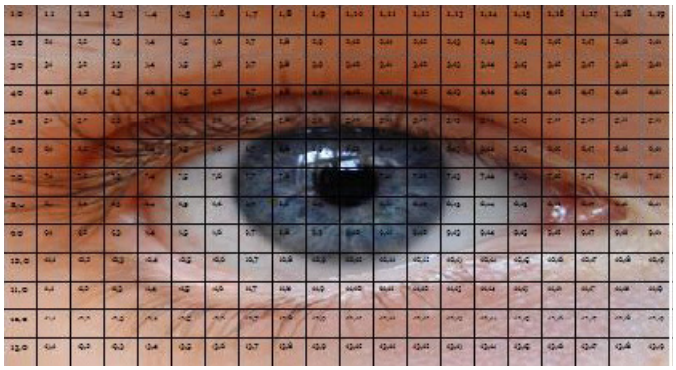


Figure 2: Captured image of the eye is scaled across the virtual grid for processing and device selection.

The grid position of the centre of cornea determines the device to be switched ON or OFF. For every grid we have one device to be operated. For the sake of simplicity, we have considered only the horizontal grids. Vertical grids could also be considered if the number of devices to be operated are more.

There are two control parameters extracted from the input video for switching the appliances. The first parameter is the centre of the cornea which determines the device. The second parameter is the motion of the eyelid i.e. Blinks. The blinks here are used as a trigger for the switch. Once we

determine the grid position of cornea, a blink can be used to switch that particular device ON or OFF. If the device is ON presently, a blink at the corresponding device grid will switch OFF the device and vice-versa.

The Blinks can be detected using another method of motion detection. For motion detection we have used frame difference algorithm for background subtraction. In this method, using live video we compare two frames of images. If the differences between the two frames is greater than a certain threshold, a motion is detected. This motion detected can be used as a trigger for switch.

The following diagram shows a real time circle detection in an image.

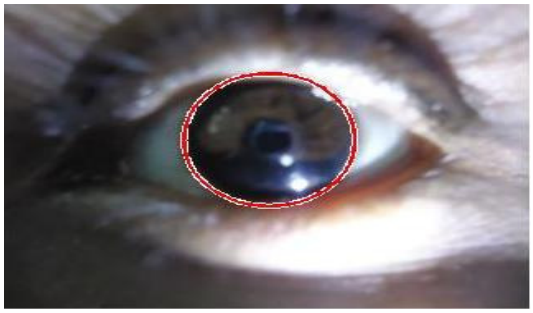


Figure 3: Circle detection using imfindcircles function of MATLAB.

The two parameters i.e. centre co-ordinate of cornea and blink output can be used in a program to give output to the hardware parts through the use of Parallel port of computer.

3. IMPLEMENTATION & RESULTS

The image processing was simulated in MATLAB and the Circular Hough Transform was used to process the image from the live camera to obtain the co-ordinates of the centre of the eye which was then used to decide which device needs to be selected. The motion detection of the eye blink was used to decide whether the device has to be switched ON or OFF. This whole process yielded a output from the parallel port of the computer which is further processed by Transmitter and Receiver blocks of the hardware.

The image processing was simulated in MATLAB and the Circular Hough Transform was used to process the image from the live camera to obtain the co-ordinates of the centre of the eye which was then used to decide which device needs to be selected. The motion detection of the eye blink was used to decide whether the device has to be switched ON or OFF. This whole process yielded a output from the parallel port of the computer which is further processed by Transmitter and Receiver blocks of the hardware.

3.1 The Transmitter Block

The Transmitter block takes the output from the parallel port of the computer and processes it with transistorized switch. The output of the transistorized switch is then fed to a DTMF(Dual tone Multiple Frequency) tone generator IC which converts the input to a DTMF output which

can be transmitted to the remote section where the devices are located. The transmitter block is as shown below:

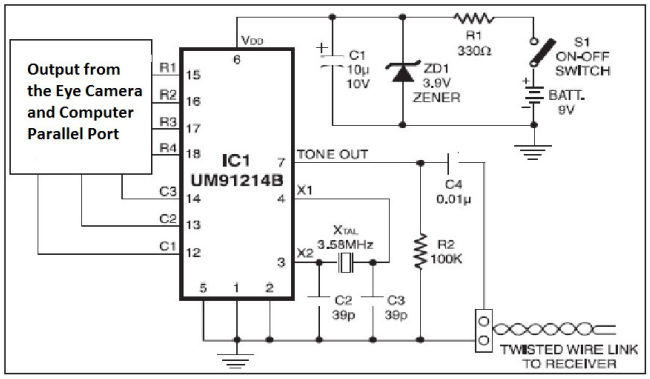


Figure 4: The Transmitter section of the Hardware part.

The transistorized switches used here are Optocouplers (not in the diagram). The optocouplers were used to enable electrical isolation between the computer parallel port and hardware section thereby preventing high voltages to affect the system.

The DTMF tone generator IC UM91214B is used to generate a single signal which can be transmitted to the Receiver section located remotely.

3.2 The Receiver Block

The receiver section consists of four parts namely, DTMF decoder, Bit Converter, Flip Flop assembly and relay circuits associated with the END devices to be operated. The circuit diagram of the Receiver block is as shown below:

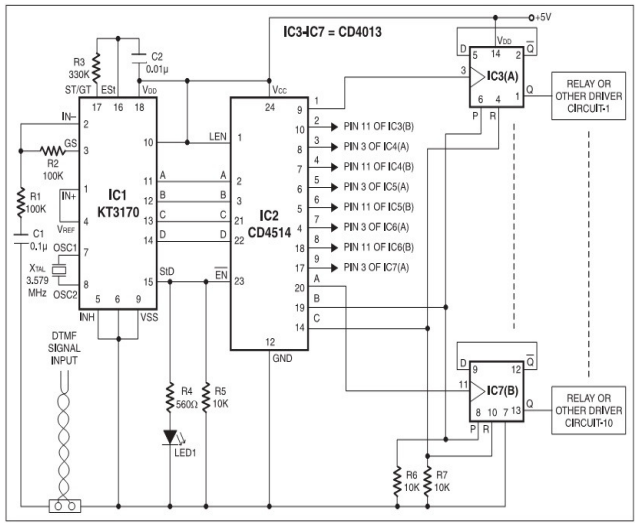


Figure 5: The Receiver section of the Hardware part.

The DTMF decoder IC KT3170 is used to decode the signal received from the transmitter section. The output of this IC is a 4-bit code containing information about which device is to be operated.

The output of DTMF decoder is fed to a 4-bit to 16-bit converter IC CD4514. The output received from this can be used to operate 16 devices individually. But, for the sake of simplicity we have used only four devices for demonstration.

The output of bit converter IC is then given to flip flop circuits. The output received from the transmitter block and parallel port may last for a very minute time. To hold this output until next triggering stage flip flops are used.

The output of the flip flops is fed to relays associated with the end appliances. The relays are used here as a switch. Since the output from the flip flops is of very low voltage, relays are used to connect the appliances to the power source. The functioning of a relay can be understood by the following diagram:

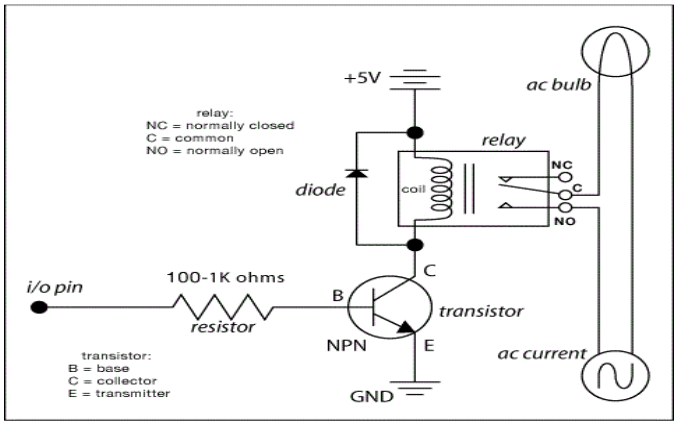


Figure 6: The Relay Assembly

The actual hardware diagram is shown below. It consists of all the above mentioned blocks.

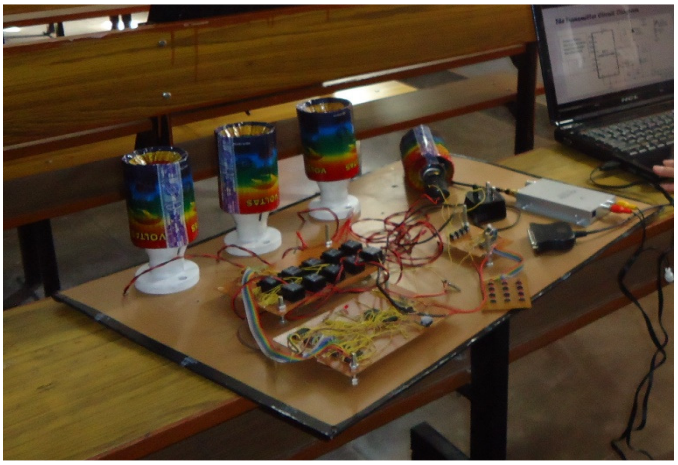


Figure 7: The hardware section.

For the sake of easiness, we tested this method for four devices which we used as AC bulb. The eye positions were scaled to the grid as shown in the below diagram:

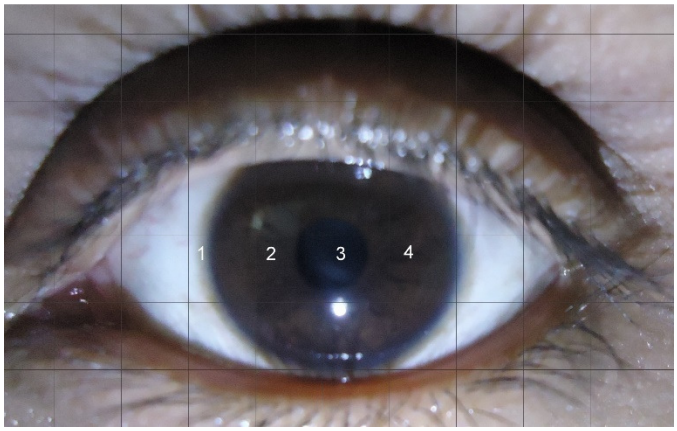


Figure 8: Scaling of eye to 4 grids for demonstration.

If the centre of the eye were to lie in grid 1, and the blinking of eye were done, the bulb 1 will get switched ON. Again if the blinking has been done while keeping the centre of the eye in grid 1, the device would turn off. Similarly, it works for all the four AC bulbs.

4. CONCLUSION

The use of standard Circular Hough Transformation in image processing module of the implementation combined with the grid analysis approach was proved to be practically successful and has high potential in future applications of the same for automated wheelchairs for the disabled. The approach has got huge potential once it is optimized in terms of the time complexity with the help of a machine with high end hardware specification. The cursor movement was achieved with a good precision in movement and the totally cost of the experimental hardware used to make the system was well below USD 150 and its commercial implementation can be easily made in affordable price range for the common man.

Moreover, the special features of the proposed method are,

1. It allows user movement. User can move in any direction when Camera is mounted.
2. It does not require any calibration process before using the proposed method.
3. It does not cause fatigue to the eyes as the headgear can be removed when not needed.

5. FUTURE WORKS

The implemented system is observed to have huge potential for use in other eye operated devices for the disabled. The future works will be aimed at modifying the approach to make it suitable for use in the eye operated wheelchairs. The wheel chair control system can be controlled using the signals generated by the system working on the discussed logic and can further reduce the cost of the expensive solutions and increase comfort for the users who are in need of the same.

ACKNOWLEDGMENT

I would like to express heartfelt gratitude towards my faculty guide Mr. Krishna Mohan Pandey, MATLAB trainer and expert, CETPA infotech Private Limited, Lucknow, or his guidance and help in completing this work successfully with good results. I would also like to express my gratitude to my Head of the Department Prof. Dr Piush Garg for his timely support and help.

REFERENCES

[1] Raj Mathews and Nidhi Chandra, Computer Mouse Tracking using Eye Tracking system based on Houghman Circle Detection Algorithm with Grid Analysis, Volume 40 No-13, International Journal of Computer Applications, February 2012.

[2] Pradipta Biswas, and Pat Langdon, A new system for disabled users involving eye gaze tracker and scanning system, Volume 5 No. 2 2011, Journal of Assistive Technologies.

[3] R.J.K. Jacob, Eye Movement-Based Human-Computer Interaction Techniques: Toward Non-Command Interfaces, in Advances in Human-Computer Interaction, Vol. 4, ed. by H.R. Hartson and D. Hix, pp. 151-190, Ablex Publishing Co., Norwood, N.J. (1993)

[4] Anjana Sharma and Pawanesh Abrol, Research issues in designing improved eye gaze based HCI techniques for Augmentative and Alternative Communication, International Journal of Emerging Technologies in Computational and Applied Sciences, 6(2), September-November, 2013, pp. 149-153.

[5] Akhil Gupta, Akash Rathi and Dr. Y Radhika, "Hands-free PC Control" controlling of mouse cursor using eye movement, Internationla journal of Scientific and Research Publications, Volume 2, Issue 4, April 2012.

[6] Kohei Arai and Ronny Mardiyanto, Eyes based Electrical Wheel Chair Control System, International Journal of Advanced Computer Science and Applications, Vol. 2, No. 12, 2011.

[7] Robert J K Jacob, The use of Eye movements in Human-Computer Interaction Techniques: What you look at is what you get, ACM Transactions on Information Systems, Vol. 9, No. 3, April 1991, Pages 152-169.

[8] Yuan-Pin Lin, 1 Yi-Ping Chao, 2 Chung-Chih Lin, and 1 Jyh-Horng Chen, Webcam Mouse Using Face and Eye Tracking in Various Illumination Environments, Proceedings of the 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference Shanghai, China, September 1-4, 2005.

[9] Web reference, Eye Controlled devices replace mouse, http://www.pcworld.com/article/47604/article.html.

[10] Web reference, Eye Tracking, Wikipedia Article, http://en.wikipedia.org/wiki/Eye_tracking.

[11] Alex Poole and Linden J. Ball, Eye tracking in Human-Computer Interaction and usability Research: Current Status and Future Prospects. Psychology Department, Lancastar University, UK

[12] Moniruzzaman Bhuiyan and Rich Picking, Gesture-controlled user interfaces, what have we done and what's next?, Center for Applied Internet Research, Glyndwr University, Wrexham, UK.

AUTHORS

Sagar Lakhmani is currently working for Accenture Services Private Limited. He has completed his graduation in Electrical and Electronics Engineering from Shri Ram Murti College of Engineering & Technology, Bareilly, U.P., India in 2013. He has worked on various projects related to Robotics and image processing. He wishes to pursue further studies on these domains.

CRYPTOGRAPHIC STEGANOGRAPHY

Vikas Yadav[1] Vaishali Ingale[2] Ashwini Sapkal[3] and Geeta Patil[4]

[1]ZS Associates, Pune, India
gunwalvikas@gmail.com
[2,3,4]Department of Information Technology, Army Institute of Technology, Pune, India
[2]vaishalidharkar@gmail.com, [3]ashwini.sapkal@gmail.com, [4]geetapatil34@gmail.com

ABSTRACT

In the cryptographic steganography system, the message will first be converted into unreadable cipher and then this cipher will be embedded into an image file. Hence this type of system will provide more security by achieving both data encoding as well as data hiding. In this paper we propose an advanced steganocryptic system that combines the features of cryptography and steganography. In this proposed steganocryptic system we will encrypt the message into cipher1 by using Kunal Secure Astro-Encryption and we again encrypt this cipher to cipher2 by using grid cipher technique. Advantage of Kunal Secure Astro-Encryption is that it generates random useless points in between, thus fixed size messages can be generated providing more security compared to other cryptographic algorithms as the number of characters in original message cannot be found from encrypted message without the knowing the black holes. Now we will embed this cipher2 into image file by using visual steganography .In this proposed steganocryptic system we will use modified bit insertion technique to achieve visual steganography. This proposed system will be more secure than cryptography or steganography techniques[digital steganography] alone and also as compared to steganography and cryptography combined systems.

KEYWORDS

kunal astro secure cryptography, grid computing, modified bit insertion technique, LSB insertion technique

1. INTRODUCTION

1.1 Overview of kunal astro secure cryptography:

In kunal astro secure cryptography method [1] we assign a unique set of coordinates, known only to receiver and sender. Another set of points known as should also be shared only between the sender and receiver.

The next step is to calculate the ASCII equivalent of each character in the message and add it with any multiple of the range of ASCII characters. Let's call this value d

Then we plot a point on the sphere with any star chosen at random from given stars and radius d. The point is plotted by taking any random value of θ and φ.

Sundarapandian et al. (Eds) : CCSEIT, DMDB, ICBB, MoWiN, AIAP - 2014
pp. 17–23, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4803

We check if this point lies inside or on any black hole or the point is already in the cipher text. If it does then we regenerate the point by going back to step involving generation of point.

We now check if this point is closest to the star from which it is generated in comparison to other stars. If not, then we regenerate the point by going back to step involving generation of point.

Now we round up the point to 2 decimal places and remove the decimal point. Now The message is further compressed by converting this number from base 10 to base 100 and append it in cipher text by separating the coordinates with a comma.

To decode it, first retrieve the triplet from cipher text each of which is separated by comma. Then we decompress it by converting from base 100 to base 10 and divide each number by 100. This will be the x, y and z values of the point.

Check if this point lies on or inside a black hole, if it does then generate nothing.

If it does not come in or on a black hole then find the star closes to this point and the distance of that star from this point. Round off this distance to nearest whole number. Take remainder by dividing this rounded distance with the maximum range of ASCII characters. This remainder is the ASCII equivalent of plain text character. Convertthis ASCII value to character to get the plain text.

The entire coordinate system has origin which is shared just between the sender and receiver.

1.2 Overview of steganography

Steganography [5] is the art and science of writing hidden messages in such a way that no one, apart from the sender and intended recipient, suspects the existence of the message, a form of security through obscurity.

There are various methods of steganography[5] e.g. Sound Technique ,Video Technique, Image hiding, Binary File Technique .In our proposed method we are going to implement Image Steganography[4][8] .The bit insertion technique is the most common technique used in image steganography ,in bit insertion technique the LSB of a pixel can be modified.

Reference [4] explains various other techniques. Many examples of LSB scheme can be found is ref [7].

1.3 Grid Cipher

It is a way to send secret messages. In grid cipher two coordinates of numbers will stand for a letter. For example, in the figure given below 68 is encrypted as ‘>‘.we can also do encryption in reverse way, like letter ‘B’ can be encrypted as ‘’86’.

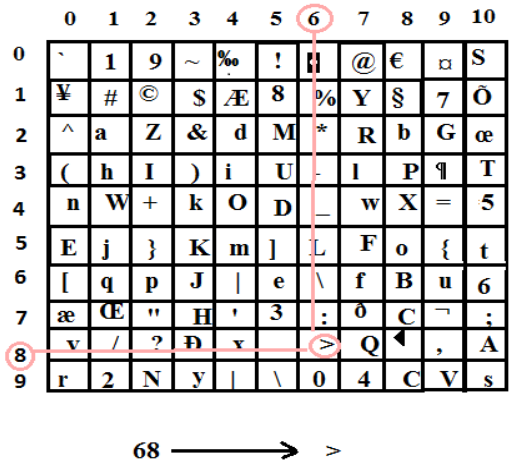


Figure 1. Grid Cipher

In our proposed system we will be having a grid database. From that grid database we will select a reference grid (this can be selected by using image properties, like depending upon width and height or it can be selected by referring size of image file or text file).Now from that reference grid(reference grid will be a collection of grids) we will select one grid (matrix) to encrypt our text.

If the total no of digits in the cipher text are odd then we will make use of our 10^{th} column of grid cipher to get cipher2 text. For example suppose our cipher text is 687 (odd number of digit) then we can decode 68 as '>' (refer figure 1) and 7(the single digit left) can be decoded as 7^{th} row of 10^{th} column i.e. ';' (refer figure 1).

Hence 687 can be decoded as '>;'.

1.4 Modified Bit insertion Technique :

In modified bit insertion technique we will first truncate the pixel values i.e. R,G,B values of a pixel to a predefined digit(for example, we can truncate the pixel values to the nearest zero digit)

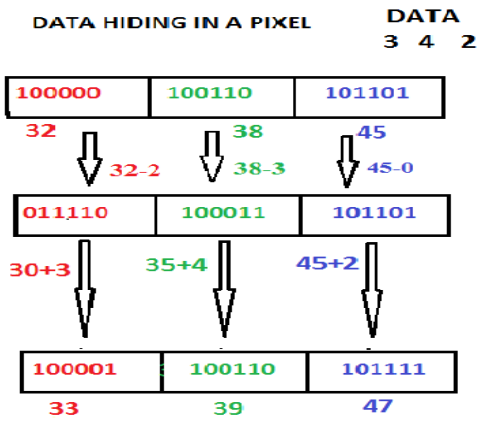


Figure 2. Modified Bit Insertion

For example, In above figure 2 first pixel value is truncated to nearest multiple of 5 and after than data is embedded. Thus the deviation in the R,G and B value of a pixel after storing the data is very less as compared to simple LSB insertion technique. Hence distortion in the image will be very less as compared to LSB insertion technique. And another advantage is that there is no need to send original image to the receiver (we have to send original image with stegano image so that receiver can decode the data from stegano image by pixel based image comparison[6] with original image).

2. THE PROPOSED METHOD

2.1 Message encryption

Step 1: We will encrypt our message by using kunal astro secure encryption into a cipher text.
Step 2: Now the cipher text obtained in step 1 is again encrypted by using matrix cipher(grid cipher technique) into cipher2 text.

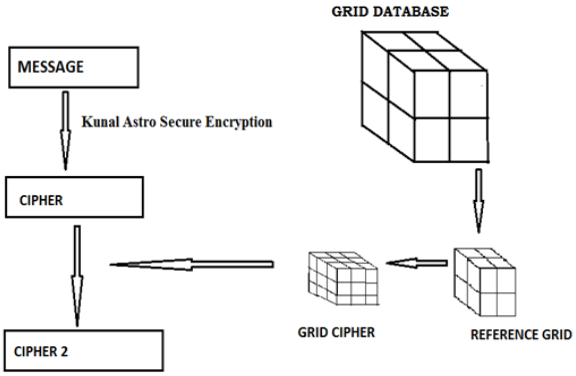


Figure 3. Message Encryption

2.2 Cipher hiding

Now we will hide the cipher2 text into an image file by using modified bit insertion technique. We will hide 1 byte of data per pixel. Here the cipher2 text file should not be too large. Amount of data that can be embed is depends on the image file. In large size file we can embed large amount of data . the cipher hiding in pixel can be understood by the figure 4.

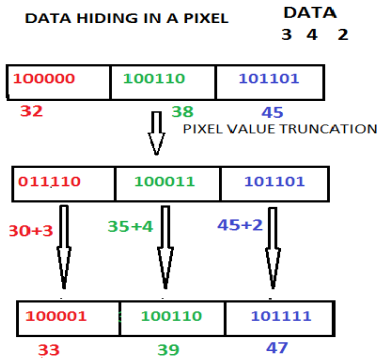


Figure 4. Cipher Hiding

2.3 Retrieval of cipher

Now we can extract the cipher2 text from our image file by the shown in the figure5 below. By doing so we will get cipher2 text.

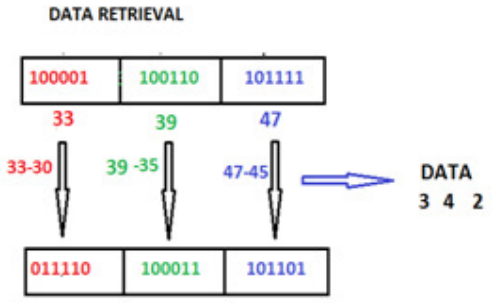


Figure 5. Retrieval of Cipher

Now just by performing reverse steps as that of encryption we can decrypt the cipher text into our plain text message. Firstly we will decrypt the cipher2 text into cipher text and now this cipher text can be decrypted to plain text message by using kunal secure astro encryption. Figure 6 shows the decryption of cipher2 into cipher text and cipher into our message.

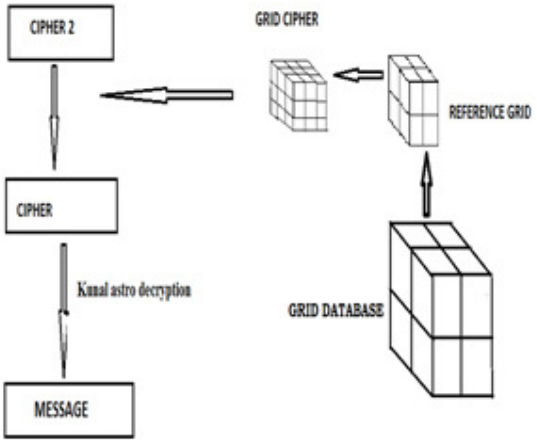


Figure 6. Decryption of cipher

3. ADVANTAGES AND APPLICATIONS

1. This method will provide more security to the information being transmitted than any other cryptographic or steganographic system as it combines both the features.

2. Extra level of security can be achieved by using grid cipher encryption

3. Distortion in the final multimedia image will be very negligible as we are using modified bit insertion technique.

The proposed system is applicable to the following areas :

1. Confidential communication and secret data storing.
2. Protection of data alteration.
3. Access control system for digital content distribution.
4. Media Database systems.

4. FUTURE SCOPE

This method can be used to increase the security on web based applications. The user will be asked to provide the secret key and the password can be compared from image files using the key. It can be used as advancement over the existing option to input the security phrase in various web based applications

REFERENCES

[1] Kush Jain, Vaishali Ingale and Ashwini Sapkal (2013) "Kunal Secure Astro-Encryption- Data Encryption and Compression Using Planar Geometry", IJETCAS.
[2] Mizuho NAKAJIMA (2002) "Extended use of Visual Cryptography for natural images, Department of Graphics and Computer Sciences", Graduate School of Arts and Sciences, The University of Tokyo.
[3] Bart Preneel (1997) "Cryptographic Algorithms: Basic concepts and application to multimedia security", Katholieke University, Belgium.
[4] T. Morkel (2005) "An Overview of Image Steganography", Department of Computer Science, University of Pretoria, South Africa.
[5] G F. Johnson and S. Jajodia (1998) "Exploring steganography: Seeing the unseen," IEEE Computer Mag., pp. 26–34.
[6] Paresh Marwaha, Piyush Marwaha and Shelly Sachdeva (2009) "Content based Image Retrieval in Multimedia Databases", International Journal of Recent Trends in Engineering .
[7] R. van Schyndel, A. Tirkel, and C. Osborne (1994) "A digital watermark," in Proc. IEEE Int. Conf. Image Processing, vol. 2, pp. 86–90.
[8] Elvin M. Pastorfide and Giovanni A. Flores (2007) "An Image Steganography Algorithm for 24-bit Color Images Using Edge-Detection Filter", Institute of Computer Science.
[9] Debashish Jena (2009) "A Novel Visual Cryptography Scheme" IEEE International Conference on Advanced Computer Control.

AUTHORS

Vikas Yadav received the BE Information Technology degree from Army Institute of Technology, Pune in 2014. he is now working in ZS Associates as a Business Technology Associate (BTA). His research interests include Security, Database

Prof Vaishali S. Ingale received the ME Computer degree from Pune University in 2008. She is now working as Assistant Professor in Army Institute Of Technology. Her research interests include Security, Algorithms, Cloud and Artificial Intelligence.

Prof. Ashwini T. Sapkal received the B.E. computer science and Engineering and Master's degrees from M.G.M's College of Engineering, Nanded, India in 2002 and Vishwakarma Institute of Technology, Pune, India in 2010, respectively. she is currently pursuing the Ph.D. degree with the Shree Guru Govind Singh Institute of Engineering and Technology, Nanded, India. She is now working as an Assistant Professor in Army Institute of Technology, Pune, India. Her current research interests include Neural Network, Pattern Classification algorithms and Cryptography.

Prof Geeta D. Patil received the ME Computer degree from College of Engineering Pune, India in 2008. She is now working as Assistant Professor in Army Institute Of Technology. Her research interests include Security Algorithms, Embedded Systems .

INTENTIONAL BLANK

USB STORAGE DEVICE CONTROL IN LINUX

Tushar B. Kute[1] and Kabita Ghosh[2]

[1 2]Department Department of Information Technology
Sandip Institute of Technology and Research Centre, Nashik
tushar@tusharkute.com,kabita18ghosh@gmail.com

ABSTRACT

The world of communication is moving towards standardization of hardware ports. All kind of communication is now using USB as the port as it is universally recognized hardware medium of communication. It is become flexible and easy to use kind of things with portable USB storage devices to copy data from one system to another system. It is possible to copy data within seconds with the help of portable USB flash memory devices. It has leaded insecurity of data storage on computer system. Various surveys has shown after network copy only USB data copy has made data insecure on computer . It is also the source of malwares in the system. To disable the USB ports is not the solution to this problem because almost all peripheral devices now uses the USB ports for communication. So, we have implemented a system which has complete USB storage enable and disable control for Linux operating system. The administrator will decide the storage devices connected to USB must be enabled or disabled .We experimented the algorithm on Linux kernel version 3.9 onwards on Debian based distributions. We have got 100% success rate of the said system with 0% performance degradation.

KEYWORDS

Linux, Debian,USB storage.

1. INTRODUCTION

With the rapid development of information technology, the communication medium has changed a lot. As communication is very important aspect of each and every work, the medium of communication has to be more efficient and more secure. Now-a-days we all had been using the USB as a port for communication for example communication between user and computers, communication between mobile and computers etc. All the peripheral hardware devices are also connected using the USB ports. So that is the reason the USB as a port are more put forward to standardisation. The most important thing that is storing of a data is also done with the USB storage devices. It makes easy to accessible to the host computing device to enable the file transfer between the two. When the USB storage devices are attached to the host computing device it appears as an external drive, to store the data. Like to copy data from one computer to another and from one computer to any storage devices. The demand for these USB storage devices has been tremendously increased. The manufacturers had also increased their production rate of these storage devices with more data storage space. But with all these flexibility, risks has also come into addition. As we are mostly using USB storage devices to transfer data or to keep backup of data, it can lead to the leakage of data. The leakage of data makes the information insecurity. This flexibility of directly accessible of copying any data from the host computing device can make the insecurity of the data. It allows the unauthorised users to access the data and copy the data from your computing device and misuse it in any ways. As now the companies in

Sundarapandian et al. (Eds) : CCSEIT, DMDB, ICBB, MoWiN, AIAP - 2014
pp. 25–29, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4804

particular are at more risk when any sensitive data are easily copied with the help of these USB storage devices by the employees and taken out of the office and misuse it or being given to any other companies. This can lead to deal with the worst consequences of losing the information that can include the customer data, business plans, financial information or some confidential documented information about company. The another risk which can lead to information loss is computer virus and other malicious software. As easily we can transfer the files between the USB storage device and the computing device at the same point of easiness the viruses can be transfer from the USB device to your computing devices. These devices has become the primary means of transmission of viruses and malware. Whenever the malware gets onto your storage device it may infect your computing devices as the USB drive is subsequently plugged. These viruses and any malicious software can corrupt your data which leads to data loss. If someone intentionally wants to corrupt all your data it just needs to plugged the storage device which contains the viruses and transfer it to your computing device. Now a days the loss of data is mainly through the computer viruses as told by the most of the surveys. For all these reasons of information insecurity, USB drives are used in a wrong manner. As information is the most valuable asset, it has to be more secure and confidential. To make the data more secure on your computers, one way is that to disable the USB ports so that no USB storage device can be plugged to your computer . But now a days almost all peripheral devices uses the USB ports for communication. So this cannot be the option to deal with the information insecurity. So in this paper a new way has been described which can be use to make our data more secure on our systems. This method has been experimented and the success rate is 100% with the 0% performance degradation. This idea has been implemented on the Linux platform which are Debian based distros and the kernel version 3.9 onwards. The idea is like only to disable the USB mass storage devices by doing some simple steps. As in Linux, only the root user has all type of authority so it can decide which user should use the USB ports for storage devices or which users should not have the privileges to plugged their mass storage devices and use it. This method can definitely improve the security of the information without losing the data and without corruption of data by any unauthorized users or by some harmful viruses.

2. LITERATURE SURVEY

As the part of academics, we take practical examination of students regularly every semester. While conduction of these examinations, we have observed that some students have tendency to copy the program from usb pen drives or through any other usb storage device. We searched the techniques to find the solution on it. We got several techniques to access and use the usb ports. These are summarized below.

The general technique to access the usb storage device [4] is with the help of commands. The command of Linux that is, usb-devices gives the listing of everything about the usb such as,

```
T:  Bus=01 Lev=00 Prnt=00 Port=00 Cnt=00 Dev#=  1 Spd=480 MxCh= 4
D:  Ver= 2.00 Cls=09(hub  ) Sub=00 Prot=00 MxPS=64 #Cfgs=  1
P:  Vendor=1d6b ProdID=0002 Rev=03.11
S:  Manufacturer=Linux 3.11.0-19-generic ehci_hcd
S:  Product=EHCI Host Controller
S:  SerialNumber=0000:00:1a.7
C:  #Ifs= 1 Cfg#= 1 Atr=e0 MxPwr=0mA
I:  If#= 0 Alt= 0 #EPs= 1 Cls=09(hub  ) Sub=00 Prot=00 Driver=hub
```

Such kind of information is made available using the command but its is difficult and almost not possible to know about usb storage devices connected to the system.

3. ARCHITECTURE

The method of disabling the USB ports for mass storage devices consists of set of Linux kernel activities that has to be included in some of the configuration files and your USB storage devices will not be detected. This method can be done only through the root account or with the help of root permission.

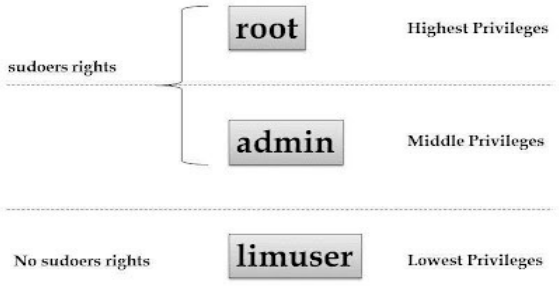


Fig.1 Account types in Linux.

The root in Linux has highest privilege rights. Administrator has middle privileges. It means in order to do any administrative activity it need the password. This is referred as sudoer rights. Sudo implies- Super User Do. The limited account user 'limuser' is not having rights to do any administrative activity in the system.
Step-1:

blacklist usb_storage

in blacklist.conf file. This configuration file reside in the /modprobe.d directory which in turn reside in the /etc directory.

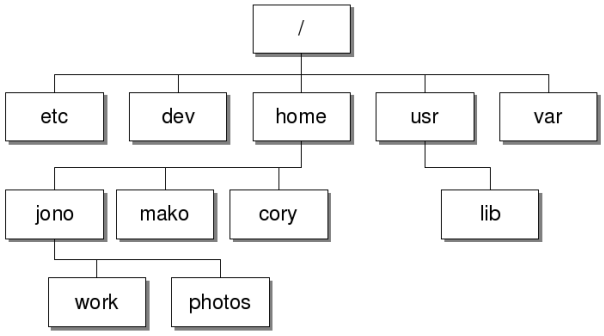


Fig.2 The Linux directory structure.

The /etc means the etc directory resides under root i.e. / represents the root. Almost all the system related configuration files are present in the /etc directory or in its sub-directories. These configuration files controls the operation of program. The modprobe.d indicates that modprobe is a directory. The modprobe is a program which adds or removes more than one module, because one module may have many dependencies, so this program is required for adding or removing multiple modules. This modprobe.d contains many .conf extension files which specifies many options that are required to be used with the modules that are to added or removed from the

kernel. The format of the .conf file underneath modprobe.d directory is very simple. It just includes one line command to configure a program and some comments regarding the command which starts from the "# ". The modprobe.d directory include a file, name blacklist.conf file which specifies the modules that has to be ignored from loading. This file helps to stop performing any operation that you don't want by not loading the modules. In blacklist.conf file, if you want to disable some operation, the keyword blacklist is used. This keyword blacklist means all of a particular module specified is ignored. In this file if you want to disable the operation, you have to use the blacklist keyword followed by the name of the module. As in this paper we are concentrating in disabling the USB storage devices so the above command i.e. blacklist keyword followed by the module name usb_storage is used. This will prevent the modprobe program for loading the USB storage driver upon demand. The usb_storage is a module related to the USB storage devices. The device drivers is the bridge between the user space applications and the hardware space. Linux kernel constantly scans all your computer buses for any changes and new hardware. Whenever the device is attached, the hardware detection is done by the USB host controller. The device signals the motherboard and the USB chip controller gets the message and says the information to the kernel with the help of an interrupt. The kernel then re-initialise the USB bus and says it to the udev that some new device is attached. The device can be detected by there identity as like all the devices have a vendor name and a model id. Then the kernel uses the modprobe program to load the driver and says the udev that there is a device of so and so vendor and model number and then the udev tries to mount the device. As discussed that the kernel invokes the modprobe program to load the drivers and modules, the modprobe program will search the configuration file whether the driver is listed or not and when it is found that the module is listed as a blacklist then the modprobe fails to load the module and the kernel could not send any information about the device and the udev could not mount the device on your system.

Step-2:

modprobe -r usb_storage

in the rc.local file which is under the /etc directory. The /etc/rc.local file is common to all major distributions. This file is empty on fresh installation and it is reserved for local administrators. The rc.local is a script file which contains any specific commands for the system that runs whenever the system startup. This file runs at the end of the system boot process, so the commands that we want to run at the time of system startup can be written in this file. The rc denotes the runcom or run command. This file can be helpful to write the commands that you want to execute at every boot time. In this file as the above command modprobe is used which helps to add or remove the modules and the dependent modules also. The kernel also depends on modprobe program to load or unload a module. There are many options that can be used with modprobe command like -a, -i etc, so the option **-r** is used in the above command which is used to remove the modules. This option is used to remove the modules and also try to remove the unused modules which are dependent on it. As whenever the storage devices are attach, the usb_storage module gets loaded which is a module related to mass storage devices and that module is used by those devices. You can see that the usb_storage module has been loaded by using the lsmod command and how many mass storage devices are attached and has used that module. This lsmod command shows the contents of the modules file under the /proc directory and the contents are the loadable kernel modules that are currently loaded on your system. So to remove that usb_storage module the modprobe program is used with -r option. To remove or unload any module you can use the modprobe program with -r option followed by the module name like usb_storage. When the system boots, at the end of all the initialization done, the rc.local file under the /etc directory gets executed and the modprobe command gets call and the usb_storage module gets unloaded.

And if we just unload the usb_storage module without including the usb_stoarge module in the blacklist.conf file then the mass storage devices get mount whenever attach because it reloads the module that had been already unloaded. So to make it persistent, the usb_storage module must be

include in the blacklist.conf file, thus restricting the module to get reload as the usb_storage module is blacklisted.

Now if we want to enable the USB mass storage devices, i.e. the storage devices to get mount we have to just remove the commands that we have added in the blacklist.conf file under the /etc/modprobe.d directory and the rc.local file under the /etc file. As the commands are removed and we attached the storage device to the system the usb_storage module get reloaded and the USB storage device get mount and can be used as normally we do use.

4. CONCLUSION

In this paper, we have proposed a system which will control the USB mass storage devices according to the administrator or root authority. The administrator can decide whether to enable or to disable the USB storage devices of the system. As a result, after disabling the USB storage devices, those devices cannot get detected and the malicious activity or malwares through these storage devices can be fully controlled by the administrator or root authority by preventing the system or confidential information to get leak or corrupt by some unauthorized users.

REFERENCES

[1] Disabling USB storage drives, March 2008, National Security Agency, USA department of defense.
[2] Defense against Malware on Removable Media, National Security Agency, USA department of defense.
[3] USB Debugging and Profiling Techniques Kishon Vijay, Abraham I, and Basak Partha, Texas Instruments, Published on http://elinux.org
[4] Robert Love; "Linux Kernel Development", 3rd Edition.
[5] Manual pages of Linux/Unix security commands..

AUTHORS

Tushar B. Kute is working as Assistant Professor in Information Technology at Sandip Institute of Technology and Research Centre, Nashik, MaharashtraHis area of interest includes the Linux Kernel Development.

Kabita K. Ghosh is student of B.E. Information Technology at Sandip Institute of Technology and Research Centre, Nashik, Maharashtra.Her area of interest includes the Linux Kernel Development.

INTENTIONAL BLANK

AUTOMATION OF ENTERPRISE AUDIT MANAGEMENT SYSTEM

Prashant P.Suryawanshi[1], Jayalaxmi G.N[2]

Department of Computer Science,
B.V.B. College of Engg & Technology,Hubli, India.
Email: [1]sssprashant11@gmail.com,[2]jaya_gn@bvb.edu

ABSTRACT

In earlier days auditing was by an independent person or body of persons with the help of vouchers, documents, information and explanations received from the authorities, for the purpose of ascertaining whether the works done entered in the books are genuine and have been entered with proper authority. It was done manually. Its job is also to find out whether they are accurate and that the works are done in accordance with law and rules and regulations of the organization in particular the standards and standard auditing practices. This drawback can be overcome using automating tool that is Automation of Enterprise Audit Tool.The Audit Tool keeps the track of the works done and intimates to higher authority if the works are not done.This application helps to manage various works in the organization And it helps to intimate the higher authority in case of the work is not done. The software "Audit Tool" helps to store the details of any enterprise and the process carried out at that enterprise.

KEYWORDS

Audit- Auditing is defined as a systematic and independent examination of data, statements, records, operations and performances (financial or otherwise) of an enterprise for a stated purpose.

1. INTRODUCTION

Internal Audit is a tool of control to measure and evaluate the effectiveness of the working of an organization primarily with accounting, financial and operational matters. The job of internal audit is to ensure that the work of the company is going on smoothly, efficiently and economically and that all the laws, rules and regulations governing the operations of the organization are adhered to, besides ensuring that an effective internal control system exists to prevent errors, frauds and misappropriations . Currently all organizations want to keep a track of all the works done at the organization. Audit tool helps in the management of the works at the organization. It helps to keep the track of the works done and intimates to higher authority if the works are not done.In earlier days auditing was by an independent person or body of persons with the help of vouchers, documents, information and explanations received from the authorities, for the purpose of ascertaining whether the works done entered in the books are genuine and have been entered with proper authority. It was done manually. Its job is also to find out whether they

Sundarapandian et al. (Eds) : CCSEIT, DMDB, ICBB, MoWiN, AIAP - 2014
pp. 31–38, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4805

are accurate and that the works are done in accordance with law and rules and regulations of the organization in particular the standards and standard auditing practices.This drawback can be overcome using the Automation of Enterprise Audit Management System.

The Advantages of Automating Enterprise Audit management are, Audit Helps To Detect And Prevent Errors And Frauds,Audit Helps To Maintain Account Regularly,Audit Helps To Get Compensation,Audit Helps To Present a Proof,Audit Provides Information About Profit Or Loss,Audit Helps To Prepare Future Plan.

The applications Automating Enterprise Audit Management System are Audit Tool is used in any Enterprise organizations, Audit Tool is used in Hospital management, Audit Tool is used in the Education Department.

The Proposed System of Automating Enterprise Audit tool works in the computerized environment and has become more relevant so as to make the audit personnel very effective in detecting irregularities. It is the examination of all managerial performance. In our audit tool system all the works done or not done can be checked online. In our system any new enterprise can registers to the system. Once the new enterprise registers to the system, the super admin has the authority to activate or delete the newly registered enterprise. The super admin also has the authority to deactivate or reactivate already registered enterprises. Once the super admin activates the newly registered enterprises a mail will be sent to that respective enterprise informing about the url, username and randomly generated password of the enterprise. The password is encrypted using MD5 encryption and stored in the database. Here the enterprise admin can add, view, edit and deletes department, stakeholder, process, checklist online. In our system as soon as the enterprise admin adds a stakeholder a mail will be sent to his/her email id informing that he has been added to so and so department with a particular role assigned to him. In our system when the enterprise admin adds a process he assigns a stakeholder to that process and as soon as the stakeholder is assigned to a process he will get a mail informing to which process he is being added. The supervisor fills the checklist online and the functional admin gets the mail of the works that have not been done along with the reason. The functional admin gets the mail continuously till the works are done. These mails are sent using crontab jobs.

The Author organizes the article in the following manner, Related Work, Proposed System, Proposed System, Conclusion and finally References.

2. RELATED WORKS

2.1 Energy Audit Tool

Author's from Green Leader's [1] and [2] about Energy Audit Tool explained as follow's. Energy audit is a crucial activity in every energy management strategy. Specific technical skills are required to efficiently perform audits of buildings. A software tool has been designed and implemented that should support every step of an exhaustive audit focused on energy usage in buildings. The software program is implemented on PDA in order to provide a portable tool that is useful for in field surveying activities. The program is structured in several procedures each focused on surveying a specific energy usage in the building. Particular attention has been given in the paper to the illustration of two particular procedures of the program that are more directly linked with electricity usage: lighting and office equipment.

2.2 Capacity Audit Tool

Author's from *Asian Development Bank (ADB). 2007a[3] and Department for International Development (DFID). 2005. [4] explained as follows about the Capacity Audit Tool.* This tool was developed out of an identified need from within the organization to create a common understanding of capacity building. On the one hand, the tool could be used by GeSCI staff in assessing its internal capacity to not only carry out its day-to-day operations but also in its capacity to offer strategic advice to its partner countries. On the other hand, the tool could be used by GeSCI while assessing the capacity of the Ministries of Education to carry out their dayto-day activities in relation to ICT4E or in executing a defined project in the same context.

2.3 District Audit Tool

Author's Abedi, J., Lord, C, Hofstetter, C., & Baker, E. (2000) [5] and Acquarelli, K., & Mumme, J. A. (1996) [6] explains the following about District Audit Tool. The No Child Left Behind (NCLB) Act of 2001 requires States to make adequate yearly progress (AYP) determinations for all districts/schools to develop school support systems for schools that do not make AYP and to provide direct support to districts in need of improvement under AYP. Given that the proportion of schools and districts likely to be identified under AYP criteria are likely to continue to increase rapidly, most States lack the capacity in fiscal and human resources to deliver uniform levels of quality support to all identified districts and to other districts in their delivery of support to schools. The support needed should be of a nature and quality that can reasonably be expected to lead to significant improvement. In addition, the same assistance may not be needed and may, in fact, interfere with the districts' progress if the assistance is inappropriate or distracting from the problems that resulted in the failure to meet AYP in the first place. Although technical assistance is provided in some form for every district missing AYP targets for two consecutive years, the intensity and the focus of assistance will vary depending on level of need and each district's specific barriers to success.

2.4 Road Safety Audit Tool

The Author's Blind Citizens Australia (2009) [7] and
Garrard, J (2013) [8] explained the following about the Road Safety Audit Tool. The document provides an accessible but comprehensive tool for assessing the walkability and safety of road environments for pedestrians with vision impairment. It is designed to be used by anyone who has an interest in road safety for people with vision impairment, from traffic engineers to volunteer advocates. It is also designed to facilitate reporting of issues in a way that gives road management authorities a comprehensive picture of the road environment as it relates to pedestrians who are blind or have low vision.information.

2.5 Manufacturing Audit to Improve Quality

Performance – A Conceptual Framework

The Author's Arter D.R. (1994) [9] and Askey, J M and Dale, B G (1994) [10] explains the following about Manufacturing Audit. Manufacturing process audit is one of the many quality tools to assess the effectiveness of manufacturing process and quality performance. They are commonly used in the effort to diagnose, maintain and improve quality management system. It is

made compulsory for the organization to maintain their quality management system based on ISO9001 standard to conduct an internal audit. However, similarly to any other physical or conceptual system, they may fail to achieve the objectives set forth, to assess effectiveness and at the same time fail to recognized area for improvement. Based on an extensive literature review, the issues relevant to manufacturing audit and quality performance are examine, and discussed the several issues to identify the conceptual framework of manufacturing audit.

3. PROPOSED SYSTEM

This project is about developing a software named "Enterprise Audit Management".This software can be used by any organization to keep the rack of all works at their organization. To use this software, the organizations should first register their organization with this software. Once the organization is activated they can start using the software. This software is an independent entity and has its own database and hence not dependant on any other softwares. This software helps to maintain the details of the stakeholders and keeps the track of works of that organization.

Architechture Diagram

This software is an independent entity and has its own database and hence not dependant on any other softwares. This software helps to maintain the details of the stakeholders and keeps the track of works of that organization

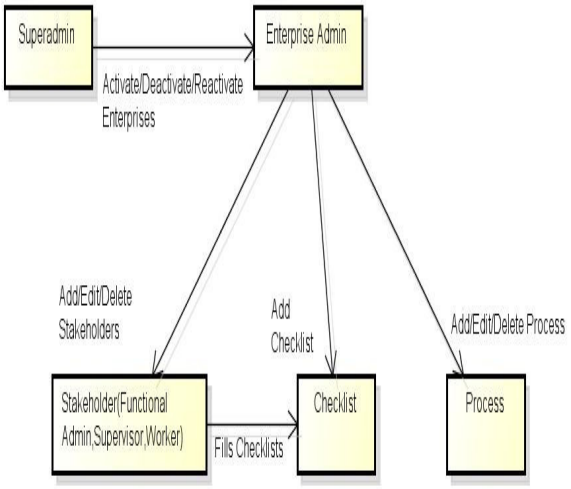


Fig.1 Architecture diagram

Super Admin:

- Manage activation, deactivation and reactivation of users (organizations).

Enterprise Admin:

- Registers his organization for the use of the audit tool.
- Creates different departments for his organization.

- View and edit the departments.
- Create stakeholder profile with role assigned to each stakeholder.
- View and edit stakeholder profile.
- Create process details and assign stakeholders to each process.
- View and edit process details.
- Create checklist for each process.
- View and edit checklist.

Functional Admin:

- Fills the checklist regarding a particular process within a given time period if not done by the supervisor.

Supervisor:

- Fills the checklist regarding a particular process within a given time period and send mail to the respective functional admin regarding the status of that process.

Algorithm 1:MD5 Encryption

MD5 is an algorithm that is used to verify data integrity through the creation of a 128-bit message digest from data input (which may be a message of any length) that is claimed to be as unique to that specific data as a fingerprint is to the specific individual. MD5, which was developed by Professor Ronald L. Rivest of MIT, is intended for use with digital signature applications, which require that large files must be compressed by a secure method before being encrypted with a secret key, under a public key cryptosystem. MD5 is currently a standard, Internet Engineering Task Force (IETF) Request for Comments (RFC) 1321. According to the standard, it is "computationally infeasible" that any two messages that have been input to the MD5 algorithm could have as the output the same message digest, or that a false message could be created through apprehension of the message digest. MD5 is the third message digest algorithm created by Rivest.

MD5 Algorithm

*Ste*p 1: *Append padded bits*

The message is padded so that the length is congruent to 448 modulo 512. A single “1” bit is appended to a message and then “0” bits are appended so that the length in bits equals 448 modulo 512.

Step 2: Append length

A 64 bit representation of b is appended to the result of the previous step. The resulting message has a length that is an exact multiple of 512 bytes.

Step 3: Initialize MD buffer

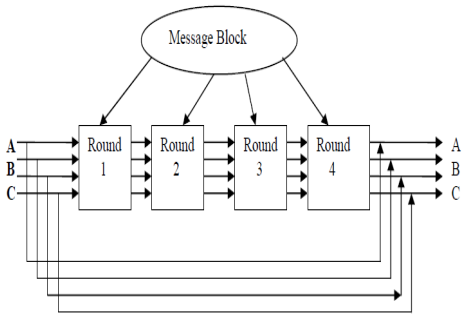
The four word buffer (A,B,C,D) is used to compute the message digest. Here each of A, B, C, D, is a 32 bit register. The registers are initialized to the following values in hexadecimal:
word A: 01 23 45 67
word B: 89 ab c def
word C: fe dc ba 98
word D: 76 54 32 10

Step 4: Process message in 64 word blocks

Four auxiliary function that take as input 3 32-bit words and produce as output 132 bit word.
F(X,Y,Z)=XY v not (X) Z
G(X,Y,Z)=XZ v Y not (Z)
H(X,Y,Z)=X xor Y xor Z
I(X,Y,Z)=Y xor (X v not (Z))
If the bits of X, Y, and Z are independent and unbiased, the each bit of F(X,Y,Z), G(X,Y,Z), H(X,Y,Z) and I(X,Y,Z) will be independent and unbiased.

Step 5: Output

The message digest produced as output is A, B, C, D, i.e., output begins with the low order byte of A and end with the high order byte of D.



Algorithm 2: Cronjobs Tab

Cron job are used to schedule commands to be executed periodically. You can setup commands or scripts, which will repeatedly run at a set time. Cron is one of the most useful tool in Linux or UNIX like operating systems. The cron service (daemon) runs in the background and constantly checks the /etc/crontab file, and /etc/cron.*/ directories. It also checks the /var/spool/cron/ directory. crontab is the command used to install, deinstall or list the tables (cron configuration file) used to drive the cron(8) daemon in Vixie Cron. Each user can have their own crontab file, and though these are files in /var/spool/cron/crontabs, they are not intended to be edited directly. You need to use crontab command for editing or setting up your own cron jobs.

Setting up a Crontab job

A crontab file consists of lines of **six** fields each.The fields are separated by spaces or tabs. The first five are integers that specify the following:
1. minute (0-59),
2. hour (0-23),
3. day of the month (1-31),
4. month of the year (1-12),
5. day of the week (0-6 with 0=Sunday).

Each of these patterns may be either an **asterisk** (meaning all valid values) or a list of elements separated by commas. An element is either a number or two numbers separated by a minus sign (meaning an inclusive range). Notice the time is in 24 hour format, **0** is midnight and **13** is one in the afternoon. The sixth field of a line in a crontab file is a string to be executed by the shell at the specified times by the first fife fields. A percent character in this field (unless escaped by \) is translated to a newline character. Only the first line (up to a % or end of line) of the command field is executed by the shell. The other lines are made available to the command as standard input. Any line beginning with a # is a comment and is ignored.

More graphically they would look like this:

```
* * * * * Command to be executed
- - - - -
| | | | |
| | | | +----- Day of week (0-7)
| | | +------- Month (1 - 12)
| | +--------- Day of month (1 - 31)
| +----------- Hour (0 - 23)
+------------- Min (0 - 59
```

4. COMPARISON

	Manual Audit	Automated Audit
Time	By an independent person or body of persons with the help of vouchers, documents, information and explanations received from the authorities, for the purpose of ascertaining the work.Hence it consumes more time	All the works done or not done can be checked online.Hence its faster than Manua audit.
Responsiveness	Response is late because a man has to report to his higher authority.	Quick response when compared to manual audit,because the automated audit works in computerized environment i e. online.
Performane	Performance is slow when compared to Automated audit because a man has to do the audit and inform to the higher authority.	Automated audit works in the computerized environment and has become more relevant so as to make the audit personnel very effective in detecting irregularities. It is the examination of all managerial performance.

5. CONCLUSION

Our audit tool works in the computerized environment and has become more relevant so as to make the audit personnel very effective in detecting irregularities. It is the examination of all managerial performance. This software can be used by any organization to keep the rack of all works at their organization. In our audit tool system all the works done or not done can be checked online.

REFERENCES

[1] http://www.phsa.ca/AboutPHSA/Environmental-Sustainability/Green-Plus-Leaders/default.html.
[2] http://www.phsa.ca/AboutPHSA/Environmental-Sustainability/Green-Plus-Leaders/default.html.
[3] Asian Development Bank (ADB). 2007a. Integrating Capacity Development into Country Programs and Operations. Medium-Term Framework and Action Plan. Manila.
[4] Department for International Development (DFID). 2005. A platform approach to Improving Public Financial Management. London.
[5] Abedi, J., Lord, C., Hofstetter, C., & Baker, E. (2000). Impact of accommodation strategies on English language learners' test performance. Educational Measurement: Issues and Practices, 19(3), 16-26.
[6] Acquarelli, K., & Mumme, J. A. (1996). Renaissance in mathematics education reform. Phi Delta Kappan, 77, 478-484.
[7] Blind Citizens Australia (2009). Policy Statement Pedestrian Safety.
[8] Garrard, J (2013). Senior Victorians and walking: obstacles and opportunities, Victoria Walks, Melbourne.
[9] Arter D.R. (1994). Quality Audit for Improved Performance. ASQC Quality Press.
[10] Askey, J M and Dale, B G (1994), Internal Quality Management Auditing:An Examination Managerial Auditing Journal, Vol. 9 No. 4, 1994, pp. 3-10, MCB University Press.

GENETIC ALGORITHM BASED HYBRID APPROACH FOR CLUSTERING TIME SERIES FINANCIAL DATA

Dr.Chandrika.J[1], Dr.B.Ramesh[2] , Dr.K.R.Ananda kumar[3] and Raina.D.Cunha[4]

[1]Dept of CS & E, M C E,Hassan, Karnataka
jc@mcehassan.ac.in
[2]Dept of CS & E, M C E, Hassan, Karnataka
br@mcehassan.ac.in
[3]Dept of CS & E, SJBIT, Bangalore, Karnataka
kra_megha_tn @hotmail.com
[4]Infosys Technology, Mysore
raina.Dcunha@gmail.com

ABSTRACT

Stock market data is a high dimensional time series financial data that poses unique computational challenges. Stock data is variable in terms of time, predicting the future trend of the prices is a challenging task. The factors that influence the predictability of stock data cannot be judged as the same factors may or may not influence the value of the stock all the time. We propose a data mining approach for the prediction of the movement of stock market. It includes using the genetic algorithm for pre processing and a hybrid clustering approach of Hierarchical clustering and Fuzzy C-Means for clustering. The genetic algorithm helps in dimensionality reduction and clustering helps to create feature vectors that help in prediction.

KEYWORDS

Time series data, Genetic Algorithm, Clustering, Stock market prediction, fuzzy C Means.

1. INTRODUCTION

Time series refers to a sequence of data points, measured typically at successive points in time spaced at uniform time intervals. Time series are used in statistics, signal processing, pattern recognition, econometrics, mathematical finance, weather forecasting, earthquake prediction, electroencephalography, control engineering, astronomy, and communications engineering. Time series analysis comprises methods for analyzing time series data in order to extract meaningful statistics and other characteristics of the data. Time series forecasting is the use of a model to predict future values based on previously observed values. Time series data have a natural temporal ordering.

Stock market prediction [1] is the act of trying to determine the future value of a company stock or a financial instrument traded on a financial exchange. Accurate predictions can help stock holders to invest further so as to gain profits from their investment, or sell their shares if there is a fall of market value. If more people want to buy a stock (demand) than sell it (supply), then the

Sundarapandian et al. (Eds) : CCSEIT, DMDB, ICBB, MoWiN, AIAP - 2014
pp. 39–52, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4806

Price moves up. Conversely, if more people wanted to sell a stock than buy it, there would be greater supply than demand, and the price would fall. Hence Stock prices change every day because of market forces such as “supply and demand”. Therefore stock data is a time series data.

Stock market prediction is one of the most challenging tasks. Many approaches and studies have been undertaken to understand the attributes that influence the stock market[1] [2]. Accurate predictions can help stock holders to invest further so as to gain profits from their investment, or sell their shares if there is a fall of market value. Since the stock market has a random behavior, accuracy of the prediction matters most for the analysts. Although, there cannot be hundred Percent accurate predictions, one can get the knowledge about the rise and fall. Stock prices change every day because of market forces such as “supply and demand”. It is easier to predict the short term price movements than the other sectors of long term market [3]. All the factors that influence the stock price are not specifically known but, to some extent the market value of short term stocks is usually influenced by structured data(price, trading volumes, accounting items) and unstructured data(financial news from newspapers, articles or internet). It is believed that it is not easy to predict how stock prices change, while certain statistical techniques on historical stock data can help to determine whether to buy or sell stock. But, stocks are volatile and can change in price rapidly [4].

Data Mining, also popularly known as Knowledge Discovery in Databases (KDD), refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. While data mining and knowledge discovery in databases (or KDD) are frequently treated as synonyms, data mining is actually part of the knowledge discovery process. Data mining automates the detection of relevant patterns in a database, using defined approaches and algorithms to look into current and historical data that can then be analyzed to predict future trends. Because data mining tools predict future trends and behaviors by reading through databases for hidden patterns, they allow organizations to make proactive, knowledge-driven decisions and answer questions that were previously too time-consuming to resolve.

The Knowledge Discovery in Databases (KDD) process is commonly defined with the following stages:

(1) Selection

(2) Pre-processing

(3) Transformation

(4) Data Mining

(5) Interpretation/Evaluation.

Selection refers to the collection of data required for the problem at hand. Data collected may be categorical data, numerical data, spatial data or temporal data. The raw data collected may be of high dimension, redundant, irrelevant, and may be prone to noise.

Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Data preprocessing is a proven method of resolving such issues. Data preprocessing prepares raw data for further processing.

Transformation of data includes dimensional reduction techniques like feature selection and feature extraction. How many features and what kind of features should be used, can be a difficult problem. There is much noise and redundancy in most high dimensionality, complex patterns. Therefore, it is sometimes difficult even for experts to determine a minimum or optimum feature set. The objective of these approaches is to find a reduced subset among the original N features such that useful class discriminatory information is included and redundant class information and/or noise is excluded.

Feature Selection is the task of finding the "best" subset of features from the initial 'N' features in the data pattern space. Feature Extraction defines a transformation from pattern space to feature space such that the new feature set used gives both better separation of pattern classes and reduces dimensionality. Thus feature extraction is a kind of feature selection, but also includes a space transformation. Feature extraction is a superset of feature selection; feature selection is a special case of feature extraction (feature extraction with the identity transformation).

Data mining uses two main core tasks clustering and classification. Clustering is an automated process to group related records together. Related records are grouped together on the basis of having similar values for attributes. This approach is an exploratory technique because it is not necessary for the end-user/analyst to specify ahead of time how records should be related together. In fact, the objective of the analysis is often to discover segments or clusters, and then examine the attributes and values that define the clusters or segments.

Classification is similar to clustering in that it also segments records into distinct segments called classes. But unlike clustering, a classification analysis requires that the end-user/analyst know ahead of time how classes are defined. For example, classes can be defined to represent the likelihood that a customer defaults on a loan (Yes/No). It is necessary that each record in the dataset used to build the classifier already have a value for the attribute used to define classes. Because each record has a value for the attribute used to define the classes, and because the end-user decides on the attribute to use, classification is much less exploratory than clustering. The objective of a classifier is not to explore the data to discover interesting segments, but rather to decide how new records should be classified.

Interpretation and Evaluation refers to the tasks of validating the results obtained through the data mining tasks.

In this paper we propose the use of both data mining techniques clustering and classification for predicting the rise or fall of stock data. The rest of the paper is organized as follows. Section 2 outlines the related work in this area. Section 3 gives a detailed account of proposed algorithm. Section 4 depicts the experimental results. Section 5 concludes the paper with direction for future enhancements.

2. RELATED WORK

Brown et. Al and Jennings et. Al [16] [17] proposed two types of stock market analysis. First, the fundamental analysis derives stock price movements from financial ratios, earnings, and management effectiveness. Second, the technical analysis identifies the trends of stock prices and trading volumes based on historical prices and volumes.

There are no specific ratios that contribute to the prediction. Various ratios calculated on the stock data are present that is derived by each analyst according to his or her observation. Hence using only limited ratios may omit an important factor that implies the most on the prediction. In the project, the historical data along with the earnings and trading volumes is considered.

Lin et al. [18] have proposed a method based on structured data such as price, trading volume and accounting items for stock market prediction. However, it is much more difficult to predict stock price movements based on unstructured textual data such as financial news published on the newspapers or Internet.

The textual data in the form of financial news and stock quotes require very high effort of fetching for some terms that are likely to be used by the companies in their documentation. After the searching is done the frequency is to be maintained and the contribution level of each word to the outcome is to be considered. This is very tedious and hard to implement. Hence historical structured data which is primarily made up of categorical data is considered.

Schumaker et.al [19] used news articles to predict stock prices. Another kind of unstructured textual data is gathered from financial reports, which contain not only textual data but also numerical data. The numerical data provides quantitative information and the textual data contains a large amount of qualitative information related to the company performance and future financial movements.

The research of Kogan et al. [20] explains that using the quantitative and qualitative information can improve the prediction accuracy. The words that are to be searched in the document are not standard and are left to the user or analyzer's choice and observations.

Though the combined effect of numerical and textual information provides more accuracy of prediction, the difficulty of implementation is observed and not included.

Genetic algorithm is used for feature selection and classification by Pie et.al [5]. Two approaches are elaborated, where Genetic algorithm is combined with the KNearest - Neighbor decision rule (GA/KNN) and a production decision rule (GA/RULE). The computational cost of the GA/KNN method was very high and required parallel or distributed processing to be attractive for large, high-dimensionality problems. As an Improvement, GA/RULE was used. The objective of the GA/RULE approach was to find an optimal transformation that yields both the lowest error and smallest feature set. The results of experiments proved that GA/RULE required substantially fewer computation cycles to achieve answers of similar quality as that of GA/KNN. The test results obtained showed that GA/RULE outperformed the standard KNN method in every case, and its performance approached the hybrid GA/KNN method in most cases. The problems faced where that the sample (training) dataset needs to be representative, and it must also be large enough to allow for effective training. Otherwise, it will allow 'false' rules to be induced.

Since stock data is of high dimension, the GA/RULE method can be used for feature selection and extraction. Since the GA/RULE works best on binary data, an approach to convert the categorical data to binary is to be considered.

Many stock prediction methods based on SVM have been proposed in [21]. The SVM-based predictive models are developed with different feature selection methods from ten years of annual reports. The results showed that document frequency threshold is efficient in reducing feature space while maintaining the same classification accuracy compared with other feature selection methods. Furthermore, the results showed the feasibility of using text classification on current year's annual reports to predict next year's company financial performance, namely the return on equity ratio.

A clustering approach for stock market prediction [6] was experimented with a hybrid approach using Hierarchical Agglomerative clustering and the K-means algorithm which was named as HRK. Both numerical and textual information from historical financial news and stock quotes were considered. The proposed method consists of three phases. First, each financial report was

converted into a feature vector and the hierarchical agglomerative clustering method was used to divide the converted feature vectors into clusters. Second, for each cluster, K-means clustering method was applied recursively to partition each cluster into sub-clusters so that most feature vectors in each sub cluster belong to the same class. Then, for each sub-cluster, the centroid was chosen as the representative feature vector. Finally, the representative feature vectors were employed to predict the stock price movements.The experimental results showed that it outperformed SVM in terms of accuracy and average profits. The total average profit of 10 industry sectors of HRK is 3.95%, while the total average profit of SVM is 1.46%. The HRK outperformed SVM to produce 73% accuracy.

From the above survey conducted, we drove to the conclusion of using only the categorical data from the historical stock data. A model is to be developed where the production rule based system (GA/RULE) can be used for finding the best rules to predict the stock price movements. The rules generated can be further combined with genetic algorithm for dimensionality reduction of the historical data considered. After the dimensional reduction, the data is given to a hybrid clustering approach that uses hierarchical agglomerative clustering and fuzzy C-Means or Hierarchical Agglomerative Clustering and K-Medoids clustering. Two hybrid approaches are used for comparative analysis. And the end result would be the accuracy of the prediction.

3. PROPOSED METHOD

The proposed work works in three phases – Data collection, Preprocessing, Dimensionality reduction and Clustering. A detailed outline of these phases is given below.

3.1 Data Collection

A company's financial data of five years is considered.[7] The five years data is divided into training data set which is of three years and testing data set which is of two years. The data should have a predefined class label for each row. '1' indicates that there is a rise and '0' indicates fall of stock value.

3.2 Pre-Processing

The training data set is pre processed to remove the row and column heading that is the date and the attribute names. Columns having empty value for each row are removed. The copy of this data is saved for future use. The mean obtained by each column is subtracted with each row of column entries. If the value obtained after subtraction is between the range +1 and -1 a value '1' is assigned if not '0' is assigned for that field. The fields are converted to 1's and 0's specifically because the data is given as an input to the genetic algorithm. The genetic algorithm works best for binary attributes. [5]

3.3 Dimensionality reduction

Genetic Algorithm is used both for classification and Feature Extraction. The production decision rule based approach is used. [5].

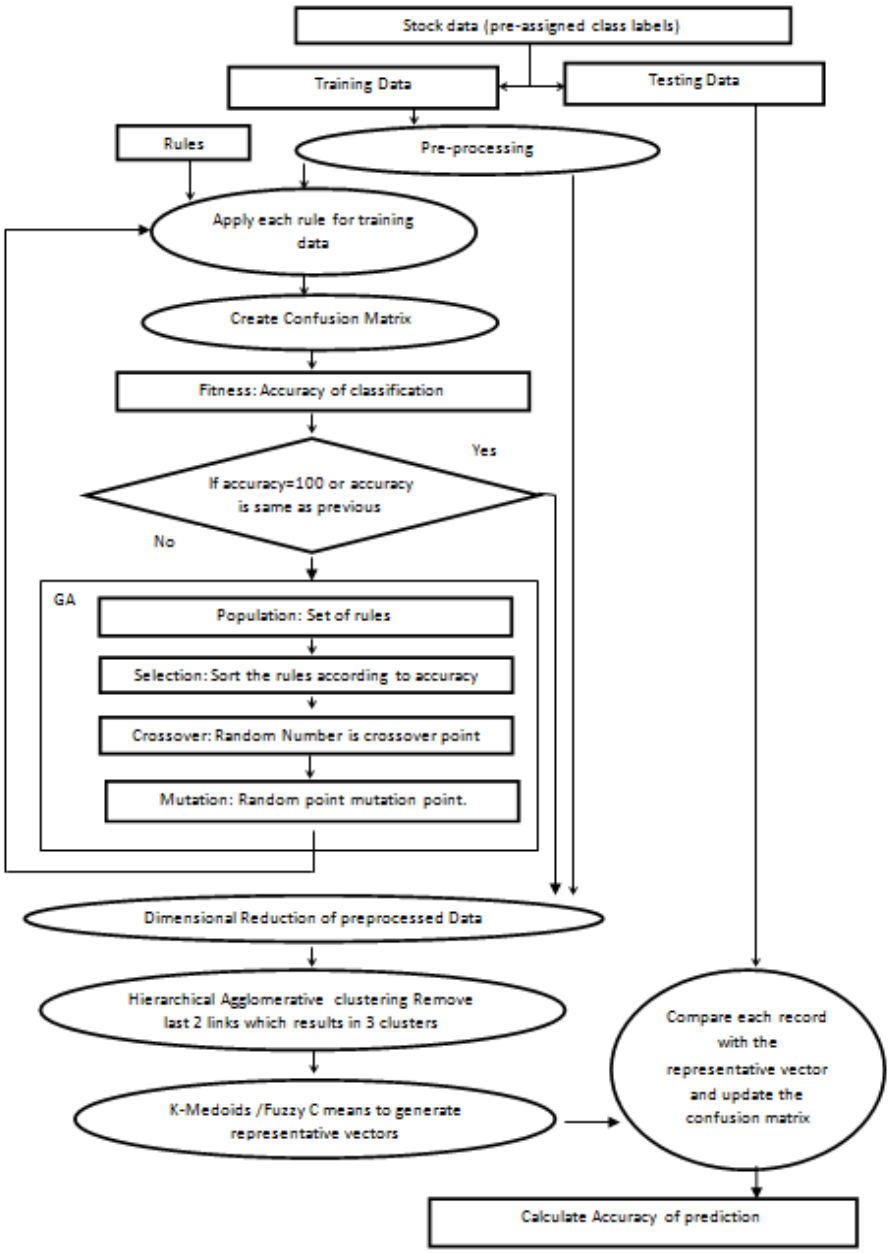


Figure 1. Outline of the model.

Genetic Algorithm (GA): Genetic algorithms imitate the evolution of the living beings, described by Charles Darwin [8]. GA is a part of the group of Evolutionary Algorithms (EA). Genetic Algorithm works with a set of individuals, representing possible solutions of the task. The selection principle is applied by using a criterion, giving an evaluation for the individual with respect to the desired solution. The best-suited individuals create the next generation. These algorithms encode a potential solution to a specific problem on a simple chromosome-like data structure and apply recombination operators to these structures as to preserve critical information. Genetic algorithms although are randomized, use historical information to find an optimal solution within the search space. The genetic algorithm is as below[9]:

1. **[Start]** Generate random population of n chromosomes (suitable solutions for the problem)
2. **[Fitness]** Evaluate the fitness f(x) of each chromosome x in the population
3. **[New population]** Create a new population by repeating following steps until the new population is complete
    a. **[Selection]** Select two parent chromosomes from a population according to their fitness (the better fitness, the bigger chance to be selected)
    b. **[Crossover]** With a crossover probability cross over the parents to form new offspring (children). If no crossover was performed, offspring is the exact copy of parents.
    c. **[Mutation]** With a mutation probability mutate new offspring at each locus (position in chromosome).
    d. **[Accepting]** Place new offspring in the new population
4. **[Replace]** Use new generated population for a further run of the algorithm
5. **[Test]** if the end conditions are satisfied, stop, and return the best solution in current population
6. [Loop] **Go to step 2**

A solution generated by genetic algorithm is called a chromosome, while collection of chromosome is referred as a population [10]. The implementation of a genetic algorithm has random chromosomes called population as input. The fitness function is applied on the chromosomes to measure the suitability of solution, more suitability gives more reproductive opportunities. Some chromosomes in population will mate through process called crossover thus producing new chromosomes named offspring which its genes composition are the combination of their parent. Mutation means random change of the value of a gene in the population. In a generation, a few chromosomes will also undergo mutation in their gene. The number of chromosomes which will undergo crossover and mutation is controlled by crossover rate and mutation rate value. Chromosomes for the next generation will be selected based on Darwinian evolution rule [11], the chromosome that achieve a better solution to the problem are given more chances to reproduce than those which are poorer. Therefore the solution is typically based on the current population. After several generations, the chromosome value will converges to a certain value which is the best solution for the problem.

Applying Genetic Algorithm - The inputs for this phase are the set of initial rules and the preprocessed data in binary. The rules have values '0' or '1' or '2' for each field and the last column of the rule indicates the class label for which the rule is defined. Then the genetic algorithm is executed with the following attributes.

Initial population: The set of initial rules represented as bit vectors.

Fitness Function: Accuracy of classification for each rule. It is based on the number of test records predicted correctly by the classification model. The frequency of incorrect and correct predictions is recorded in a table called confusion matrix.

If f_{ij} indicates the number of records of class 'i' predicted as the records of class 'j', then f_{11} and f_{00} are the number of correct predictions while f_{10} and f_{01} are the number of incorrect predictions.

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions}$$

$$Accuracy = \frac{f11 + f00}{f00 + f11 + f01 + f10}$$

A classification model should always try to attain higher accuracy.

Selection: The genetic algorithm arranges the rules based on this fitness function i.e. the rule having the highest accuracy is at the top. Then the genetic algorithm selects the top two rules to perform crossover and mutation.

Crossover: Crossover takes place on the selected rules at a randomly generated point.

Mutation: The mutation rate considered is 2%. The mutation occurs at a randomly generated point.

Population: The rules generated after crossover and mutation are the population for the next iteration of the genetic algorithm.

Stopping Condition: The algorithm is executed until, the accuracy reaches 100 or the genetic algorithm executes for 1000 generations.

The genetic algorithm executes repeatedly to generate the best set of rules for predicting stock market data. The rules that are generated by the genetic algorithm are analyzed. If a column has a value 2 for each rule that column in the saved training data is deleted. Because a value 2 indicates that the value in that field doesn't contribute to the outcome. Hence we get a data set with reduced dimensionality.

3.4 . Clustering

The dimensional reduced data is given for the clustering process. We propose a hybrid clustering mechanism, which includes Hierarchical Agglomerative Clustering and Fuzzy-C-Means Clustering or Hierarchical Agglomerative Clustering and K-Medoids clustering. A comparative analysis is to be performed using Fuzzy C-Means and K-medoids.

1) *Hierarchical Agglomerative Clustering:* Hierarchical agglomerative clustering [12] or HAC is a bottom-up hierarchical clustering technique. Bottom-up algorithms treat each document as a singleton cluster at the outset and then successively merge (or agglomerate) pairs of clusters until all clusters have been merged into a single cluster that contains all documents. An HAC clustering is typically visualized as a dendrogram. Each merge is represented by a horizontal line. The y-coordinate of the horizontal line is the similarity of the two clusters that were merged, where documents are viewed as singleton clusters. Hierarchical clustering does not require a pre-specified number of clusters.

1. Start by assigning each item to a cluster, so that if you have N items, you now have N clusters, each containing just one item. Let the distances (similarities) between the clusters the same as the distances (similarities) between the items they contain.
2. Find the closest (most similar) pair of clusters and merge them into a single cluster, so that now you have one cluster less.
3. Compute distances (similarities) between the new cluster and each of the old clusters.
4. Repeat steps 2 and 3 until all items are clustered into a single cluster of size N. (*)

Step 3 can be done in different ways, single-linkage, complete-linkage and average-linkage clustering.

In single-linkage clustering (also called the connectedness or minimum method), we consider the distance between one cluster and another cluster to be equal to the shortest distance from any member of one cluster to any member of the other cluster. If the data consist of similarities, we

consider the similarity between one cluster and another cluster to be equal to the greatest similarity from any member of one cluster to any member of the other cluster.

In complete-linkage clustering (also called the diameter or maximum method), we consider the distance between one cluster and another cluster to be equal to the greatest distance from any member of one cluster to any member of the other cluster.

In average-linkage clustering, we consider the distance between one cluster and another cluster to be equal to the average distance from any member of one cluster to any member of the other cluster.

Fuzzy C-Means*:* FCM algorithm is one of the most important fuzzy clustering methods, initially proposed by Dunn, and then generalized by Bezdek [13]. The Fuzzy C-means clustering algorithm is a variation of the K-means clustering algorithm, in which a degree of membership of clusters is incorporated for each data point. FCM algorithm is a technique of clustering which permits one piece of data to belong to two or more clusters [14]. The aim of the FCM algorithms is to assign the data points into clusters with varying degrees of membership values. Membership values lie between 0 and 1. This membership value reflects the degree to which the point is more representative of one cluster than the other. The centroids of the clusters are computed based on the degree of memberships as well as data points. The algorithm consists of the following steps:

1. Let us suppose that M-dimensional N data points represented by x_i (i = 1, 2. . . N) are to be clustered.
2. Assume the number of clusters to be made, that is, C, where $2 \leq C \leq N$.
3. Choose an appropriate level of cluster fuzziness f >1.
4. Initialize the N × C × M sized membership matrix U, at random, such that $U_{ijm} \in [0,1]$ and $\sum_{j=1}^{c} U_{ijm}$=1.0, for each i and a fixed value of m.
5. Determine the cluster centers CC_{jm}, for j cluster and its m dimension by using the expression given below:

$$CC_{\text{jm}}=\frac{\sum_{i=1}^{N} U_{ijm}^{f} x_{im}}{\sum_{i=1}^{N} U_{ijm}^{f}}$$

6. Calculate the Euclidean distance between i^{th} data point and j^{th} cluster center with respect to say m^{th} dimension like the following:

$$D_{ijm}=\left\|\left(x_{im}-CC_{jm}\right)\right\|$$

7. Update fuzzy membership matrix U according to D_{ijm}. If $D_{ijm} > 0$, then

$$CC_{ijm}=\frac{1}{\sum_{c=1}^{C}\left(\frac{D_{ijm}}{D_{icm}}\right)^{\frac{2}{f-1}}}$$

   If D_{ijm}= 0, then the data point coincides with the corresponding data point of j^{th} cluster center C_{jm} and it has the full membership value, that is, U_{ijm}= 1.0.
8. Repeat from Step 5 to Step 7 until the changes in the U ≤ ∈, where ∈ is a pre-specified termination criterion.

K-Medoids Clustering: The k-medoids algorithm is a clustering algorithm related to the k-means algorithm and the medoidshift algorithm [15]. K-medoid is a classical partitioning technique of clustering that clusters the data set of n objects into k clusters known a priori. It is more robust to

noise and outliers as compared to k-means because it minimizes a sum of pair wise dissimilarities instead of a sum of squared Euclidean distances.

A medoid can be defined as the object of a cluster, whose average dissimilarity to all the objects in the cluster is minimal i.e. it is a most centrally located point in the cluster. The algorithm can be given as:

1. Initialize: randomly select (without replacement) k of the n data points as the medoids
2. Associate each data point to the closest medoid.
   ("closest" here is defined using any valid distance metric, most commonly Euclidean distance, Manhattan distance or Minkowski distance)
3. For each medoid m
4. For each non-medoid data point o
   a. Swap m and o and compute the total cost of the configuration
5. Select the configuration with the lowest cost.
6. Repeat steps 2 to 4 until there is no change in the medoid.

The dimensional reduced data is given to the hierarchical clustering algorithm. The last two links are removed to generate three clusters of the data. The Euclidean distance is used as a proximity measure. It is applied to two data points which lie in 1, 2, 3 or higher dimensional space. It is calculated as

$$d(x,y) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2}$$

Where, n is the number of dimensions, x_k is the value of the data object x for dimension k, y_k is the value of the data object y for dimension k.

Each cluster is then given to the Fuzzy C-Means or the K-medoids algorithm to perform sub clustering. The Euclidean distance is used as a proximity measure. The clustering is performed until each cluster has a purity of 1. The centroid of each cluster along with the class label is transformed to a representative vector.

3.5 Prediction

The testing data is compared to each representative vector to find to which centroid the testing record is the closest. Then the class label of the representative vector is used as the predicted value for the testing data. This is done for each testing record. Since the testing data has a pre assigned class label, the accuracy of prediction is calculated.

4. EXPERIMENTATION

The proposed algorithm is implemented in MATLAB R2009b version. It provides plenty of user interface controls for creating a dynamic User Interface. It has several inbuilt functions that aids in the efficient implementation of proposed algorithm. The training data set, the set of rules and the testing data set are all stored as files. The software provides the flexibility of reading and displaying the file with numerical data or strings (rules).The training data is stored as a matrix format. The Rules are stored in a list and each rule is considered as a string. The software provides various internal functions for data type conversions and storing the data. For pre processing the column headers and the first column is deleted just by specifying the row and column index. To check for empty values an inbuilt function ‘isnan’ is used. The ‘fcm’ and

'kmeans' functions are already inbuilt that correspond to Fuzzy C-Means and K-Means (if the attributes are greater than 2 it works as K-Medoids) algorithms. The representative vectors and the testing data are stored in matrix format. The accuracy obtained by the genetic algorithm for feature extraction is 87%. The purity of each sub cluster obtained is 1. The accuracy of prediction obtained is 88%. Hence the model developed is better than the models developed previously.

Figure 2 shows the percentage of feature reduction obtained through the use of Genetic algorithm. The reduction rate of feature extraction obtained is 40%.

Since the accuracy of prediction obtained by both the algorithms is same, it is essential to observe the execution time taken by the hybrid algorithm using both the clustering approaches that is K-Medoids and Fuzzy C-Means. Figure 3 shows that Fuzzy C-Means takes more time to execute than K-Mediods as the number of clusters increase.

The accuracy obtained by our model is compared to the accuracy obtained from the existing model. From Fig 5 and Fig 4 we see that the HRK approach has achieved an accuracy of 73% while HAC with K-Medoids or HAC with Fuzzy C-Means has achieved 88% accuracy.

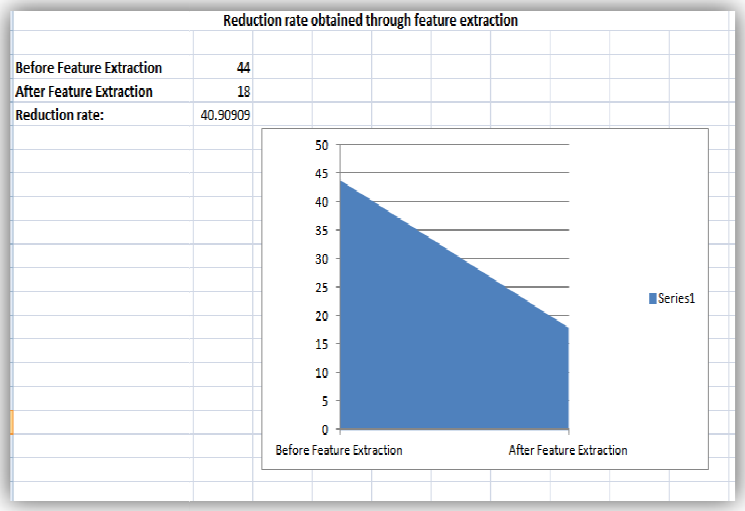


Figure 2. Reduction rate through Feature Extraction

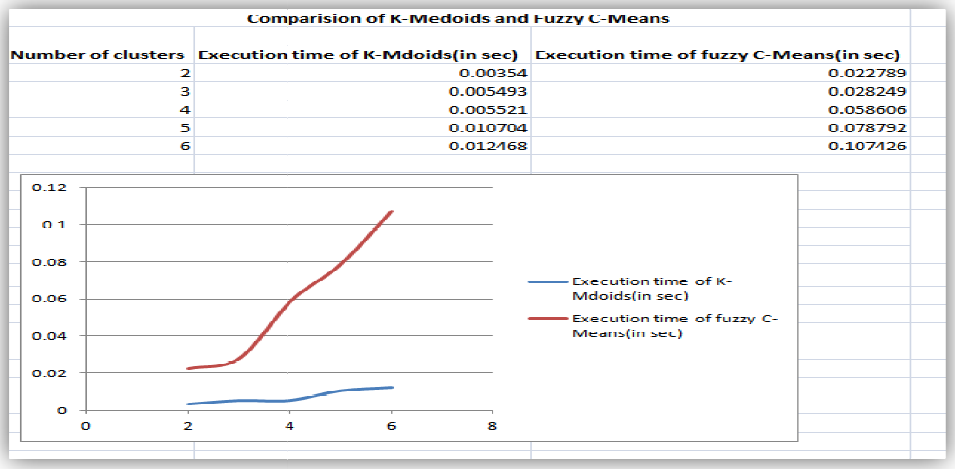


Figure 3. Comparison of fuzzy C-Means and K-Medoids

Figure 4. Accuracy achieved by existing Algorithms

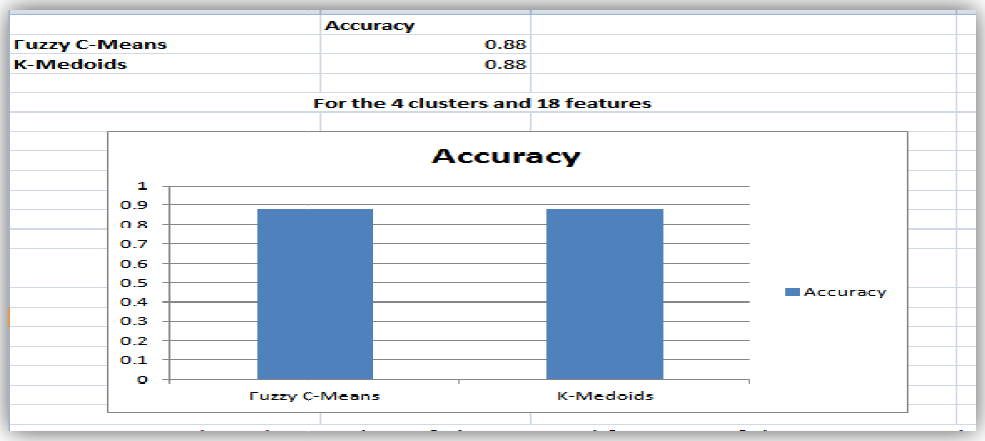


Figure 5 Accuracy obtained by proposed algorithms

5. CONCLUSION AND FUTURE ENHANCEMENTS

Data mining is found to be a useful domain that can be used to predict the stock price value and to build a structural model to predict accurate results. An efficient prediction can help the investors to gain a huge amount of profit. It is experimented that combining the methods of preprocessing, classification and clustering gives more accurate results than the other methods proposed to date. The model developed is highly dependent on the initial rules used. Even though the genetic algorithm is used for efficient processing of the rules, if the initial rules do not even provide half of the accuracy of classification, the genetic algorithm will evolve generating inefficient algorithms for feature extraction. From the experimental results conducted we found that fuzzy C-Means and K-Medoids both achieved an accuracy of 88% when using the same data set, rule set and the same number of clusters. But the execution time Fuzzy C-Means is greater than that of K-Medoids. Hence we conclude that K-Medoids is a better clustering algorithm for the model developed.

The future enhancements to this work would be considering categorical data along with numerical data. Instead of building a model for a particular company, it would be efficient to construct a general model so that even long-term company price movements can be predicted.

REFERENCES

[1] S Abdulsalam Sulaiman Olaniyi, Adewole, Kayode S., Jimoh, R. G (July 2011), "Stock Trend Prediction Using Regression Analysis –A Data Mining Approach", ISSN 2222-9833 ARPN Journal of Systems and Software.

[2] Clive W.J. G-anger, (1992),"Forecasting stock market prices: Lessons for forecasters", International Journal of Forecasting .

[3] http://www.indianstocktimes.com/study-zone.php

[4] Hongxing He,Jie Chen,Huidong Jin,Shuheng Chen(2006),"Stock Trend Analysis and Trading Strategy", ,jcis ,Taiwan 2006, DOI: 10.2991/jcis.2006.

[5] Min Pei, Erik D. Goodman, William F. Punch III and Ying Ding, (1995),"Genetic Algorithms For Classification and Feature Extraction", Michigan State University, Genetic Algorithms Research and Applications Group (GARAGe),CSNA-95.

[6] M.Suresh Babu, Dr. N.Geethanjali, Prof B.Satyanarayana,, (2012), "Clustering Approach to Stock Market Prediction", Int. J. Advanced Networking and Applications Volume: 03, Issue: 04, Pages:1281-1291.

[7] http://finance.yahoo.com/q/hp?s=YHOO

[8] http://www.ro.feri.unimb.si/predmeti/int_reg/Predavanja/Eng/3.Genetic%20algorithm/_25.html

[9] http://en.wikipedia.org/wiki/Genetic_algorithm.

[10] Tom V. Mathew, "Genetic Algorithm", Indian Institute of Technology Bombay, Mumbai, available at http://www.civil.iitb.ac.in/tvm/2701_dga/2701-ga-notes/gadoc.pdf

[11] Ganesh Bonde, Rasheed Khaled ,(2012),"Stock price prediction using genetic algorithms and evolution Strategies", ,Intl. conference on artificial intelligence(ICAI2012), Las vegas.

[12] http://nlp.stanford.edu/IR-book/html/htmledition/hierarchical-agglomerative-clustering-1.html

[13] N.R, Pal K, Keller J.M. and Bezdek J.C, (2005),"A Possibilistic Fuzzy c-Means Clustering Algorithm", IEEE Transactions on Fuzzy Systems, Vol. 13, No. 4, Pp. 517–530.

[14] A.vathy-Fogarassy, B.Feil, J.Abonyi (2005),"Minimal Spanning Tree based Fuzzy clustering", Proceedings of World academy of Sc., Eng & Technology, vol-8.

[15] http://en.wikipedia.org/wiki/K-medoids

[16] David.P.Brown, Robert.H.Jennings (1989),"On Technical Analysis, in Review of financial studies", vol.2, issue 4,pp 527-557.

[17] Jeffery.S.Ababanell and Brain.J Bushee, (1998),"Abnormal returns to fundamental analysis strategy", The Accounting Review, Vol.73.

[18] ",L.Lin,Ren.R.E,D.Sornette,"Consistent model of explosive financial bubbles with mean reversing residuals", arXiv:0905.0128

[19] Robert.P.schumaker,Hisinchun chen,(2009), "Textual analysis of stock market prediction using breaking financial news: The AZFin text system, ACM transactions on information systems,vol. 27, issue 2.

[20] Shimon Kogan,Dimitry Levin,Bryan R.Routledge, Jacob.S.Sagi,Noah.A.Smith,(2009), "Predicting risks from financial reports with regression", Available at http://svmlight.joachims.org.

[21] TAY, Francis E. H. and Lijuan CAO, (2001),"Application of support vector machines in financial time series forecasting", Omega: The International Journal of Management Science, Volume 29, Issue 4, Pages 309-317

[22] Yuling LIN, Haixiang GUO and Jinglu HU, (2013),"An SVM-based Approach for Stock Market Trend Prediction", Proceedings of International Joint Conference on Neural Networks, Dallas, Texas, USA.

AUTHORS

Dr. J.Chandrika holds doctoral degree in computer science and engineering. Currently works as Associate professor in the department of computer science and engineering at Malnad college of Engineering, Hassan,Karnataka .Her areas of interest include, data stream mining ,Artificial intelligence and medical data mining She has six international conference publications and four international journal publications and one national conference publication to her credit.

Dr.B.Ramesh holds a doctoral degree in computer science and engineering. Currently works as Professor and Head of the department in the department of computer science and engineering ,MCE Hassan.He has a vast teaching experience of about 21 years.His research interest includes mobile adhoc networks, Computer networks and Data mining. He is currently guiding five research scholars.

Dr. K.R.Ananda kumar holds a doctoral degree in computer science and engineering. Currently works as Professor and Head of the department in the department of computer science and engineering ,SJBIT Bangalore.He has a vast teaching experience of about 25 years.His research interest includes medical data mining, data stream mining ,Artificial intelligence, Intelligent agents and web mining. .He has successfully guided two doctoral candidates. He has many publications in nternational and national journals to his credit.

Raina.D.cunha is a B.E student at Malnad College of Engineering Hassanand will be graduating in the year 2014. Presently she is working as trainee systems engineer at Infosys technology Mysore. She has undertaken many Data Mining projects during her UG course.

EARLY HEART DISEASE PREDICTION USING DATA MINING TECHNIQUES

Aditya Methaila[1], Prince Kansal[2], Himanshu Arya[3], Pankaj Kumar[4]

Student, B.Tech (CSE),Maharaja Surajmal Institute of Technology
New Delhi, India

ABSTRACT

The successful application of data mining in highly visible fields like e-business, marketing and retail has led to its application in other industries and sectors. Among these sectors just discovering is healthcare. The Healthcare industry is generally "information rich", but unfortunately not all the data are mined which is required for discovering hidden patterns & effective decision making .Discovery of hidden patterns and relationships often goes unexploited. Advanced data mining modeling techniques can help remedy this situation.

This research paper intends to use data mining Classification Modeling Techniques, namely, Decision Trees, Naïve Bayes and Neural Network, along with weighted association Apriori algorithm and MAFIA algorithm in Heart Disease Prediction. Using medical profiles such as age, sex, blood pressure and blood sugar it can predict the likelihood of patients getting heart disease.

1. INTRODUCTION

Data mining is the process of finding previously unknown patterns and trends in databases and using that information to build predictive models. Data mining combines statistical analysis, machine learning and database technology to extract hidden patterns and relationships from large databases.

The World Health Statistics 2012 report enlightens the fact that one in three adults worldwide has raised blood pressure – a condition that causes around half of all deaths from stroke and heart disease. Heart disease, also known as cardiovascular disease (CVD), encloses a number of conditions that influence the heart – not just heart attacks. Heart disease was the major cause of casualties in the different countries including India. Heart disease kills one person every 34 seconds in the United States. Coronary heart disease, Cardiomyopathy and Cardiovascular disease are some categories of heart diseases. The term "cardiovascular disease" includes a wide range of conditions that affect the heart and the blood vessels and the manner in which blood is pumped and circulated through the body. Diagnosis is complicated and important task that needs to be executed accurately and efficiently. The diagnosis is often made, based on doctor's experience & knowledge. This leads to unwanted results & excessive medical costs of treatments provided to patients. Therefore, an automatic medical diagnosis system would be exceedingly beneficial. Our work attempts to present the detailed study about the different data mining techniques which can be deployed in these automated systems.

Sundarapandian et al. (Eds) : CCSEIT, DMDB, ICBB, MoWiN, AIAP - 2014
pp. 53–59, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4807

2. METHODOLOGY

This paper exhibits the analysis of various data mining techniques which can be helpful for medical analysts or practitioners for accurate heart disease diagnosis. Due to resource constraints and the nature of the paper itself, the main methodology used for this paper was through the survey of journals and publications in the fields of medicine, computer science and engineering.

3. RESEARCH FINDINGS

3.1 Data Mining in the Heart Disease Prediction.

Different supervised machine learning algorithms i.e. Naïve Bayes, Neural Network, along with weighted association Apriori algorithm, Decision algorithm have been used for analyzing the dataset in [1]. The data mining tool Weka 3.6.6 is used for experiment. Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes.

3.1.1 Decision Tree is a popular classifier which is simple and easy to implement. There is no requirement of domain knowledge or parameter setting and can high dimensional data can be handled. It produces results which are easier to read and interpret. The drill through feature to access detailed patients' profiles is only available in Decision Trees.

3.1.2 Naïve Bayes is a statistical classifier which assumes no dependency between attributes. This classifier algorithm uses conditional independence, means it assumes that an attribute value on a given class is independent of the values of other attributes. The advantage of using naïve bayes is that one can work with the Naïve Bayes model without using any Bayesian methods.

3.1.3 Neural Networks

An artificial neural network (ANN), often just called a "neural network" (NN), is a mathematical model or computational model based on biological neural network. In other words, it is an emulation of biological neural system [9]. In feed-forward neural networks the neurons of the first layer forward their output to the neurons of the second layer, in a unidirectional fashion, which explains that the neurons are not received from the reverse direction. There is connection between each layer & weights are assigned to each connection. The primary function of neurons of input layer is to divide input xi into neurons in hidden layer. Neuron of hidden layer adds input signal xi with weights wji of respective connections from input layer. The output Yj is function of

$$Yj = f(\Sigma\ wji\ xi)$$

where f is a simple threshold function such as sigmoid or hyperbolic tangent function.

4. DATA SOURCE

Using medical profiles such as age, sex, blood pressure and blood sugar it can predict the likelihood of patients getting a heart disease. It enables significant knowledge, e.g. patterns, relationships between medical factors related to heart disease. The publicly available heart disease database is used which can be used for determining various heart diseases.

Input Attributes:

1. age: age in years
2. sex: sex (1 = male; 0 = female)
3. cp: chest pain type
Value 1: typical angina
Value 2: atypical angina
Value 3: non-anginal pain
Value 4: asymptomatic
4. trestbps: resting blood pressure (in mm Hg on admission to the hospital)
5. chol: serum cholestoral in mg/dl
6. fbs: (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
7. restecg: resting electrocardiographic results Value 0: normal
Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
8. thalach: maximum heart rate achieved
9. exang: exercise induced angina (1 = yes; 0 = no)
10. oldpeak = ST depression induced by exercise relative to rest
11. slope: the slope of the peak exercise ST segment Value 1: upsloping
Value 2: flat
Value 3: downsloping
12. ca: number of major vessels (0-3) colored by flourosopy
13. thal: 3 = normal; 6 = fixed defect; 7 = reversable defect
14. num: diagnosis of heart disease (angiographic disease status)
Value 0: < 50% diameter narrowing
Value 1: > 50% diameter narrowing

Key Attributes

1. Patientid – Patient's identification number

A total of 909 records were obtained from the Cleveland Heart Disease database. It has been observed during the analysis that Naive Bayes appears to be most effective as it has the highest percentage of correct predictions (86.53%) for patients with heart disease, followed by Neural Network (85.53%) and Decision Trees. Decision Trees, however, appears to be most effective in case of predicting patients with no heart disease,

i.e. (89%) as compared to other two models.

Techniques	Accuracy
Naive Bayes	86.53 %
Decision Tree	**89%**
ANN	85.53

In the research the number of attributes which were used for heart disease diagnosis were reduced .Earlier,13 attributes were used for this prediction but this research work reduced the number of attributes to six only using Genetic Algorithm and Feature Subset Selection.

Genetic Algorithm [6] incorporates natural evolution methodology. The genetic search starts with zero attributes, and an initial population with randomly generated rules .Based on the idea of survival of the fittest, new population was constructed to match with fittest rules in the current population, as well as offspring of these rules. Offspring were generated by applying genetic operators; cross over and mutation. The process of generation continued until it evolved a population P where every rule in P satisfied the fitness threshold. With initial population of 20 instances, generation continued till the twentieth generation with cross over probability of 0.6 and mutation probability of 0.033. The genetic search resulted in six attributes out of thirteen attributes .After reduction of 13 attributes to 6 attributes, various classifiers are used on the dataset corresponding to these 6 attributes for heart disease prediction.

Performance analysis of these classifiers is shown in Table 4. It can be perceived from the table that Decision Tree has outperformed with highest accuracy and least mean absolute error.

Techniques	Accuracy
Naive Bayes	96.53 %
Decision Tree	99.2%
Classification via clustering	88.3

5. ASSOCIATIVE CLASSIFICATION

Associative classification mining is a promising approach in data mining that utilizes the association rule discovery techniques to construct classification systems, also known as associative classifiers. . Association rule mining is used to find associations or correlations among the item sets. It is a unsupervised learning where no class attribute is involved in finding the association rule. On the other hand, classification is a supervised learning where class attribute is involved in the construction of the classifier and is used to classify or predict the data unknown sample. Associative classification is a recent and rewarding technique which integrates association rule mining and classification to a model for prediction and achieves maximum accuracy. Associative classifiers are especially fit to applications where maximum accuracy is desired to a model for prediction. Various Techniques which can be used are apriori algorithm, éclat algorithm, FP-growth algorithm. Here we will be using Apriori Algorithm for discovering interesting relations in heart based diseases.

Apriori Algorithm:

Apriori is an algorithm for frequent item set mining and association rule learning over transactional databases. It proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database. A frequent itemset can be a defined as a subset of frequent itemset i.e., if {PQ} is a frequent itemset, both {P} and {Q} should be a frequent itemset.

1. Iteratively discover frequent itemsets with cardinality from 1 to k (k-itemset).
2. Use the frequent itemsets produce association rules. Join Step: Ck is generated by joining Lk-1with itself Prune Step: Any (k-1)-itemset that is not frequent cannot be a subset of a frequent k-itemset Initialize: K: = 1, C1 = all the 1- item sets; read the database to count the support of C1 to determine L1. L1 := {frequent 1- item sets};
k: =2;
//represents the pass number//
While (Lk-1 ≠) do
begin
Ck: = gen_candidate_itemsets with the given Lk-1 Prune (Ck) for all candidates in Ck do
count the number of transactions of at least k length that are common in each item Ck Lk := All candidates in Ck with minimum support;
k := k + 1;
end

Frequent Pattern mining using MAFIA

Mining frequent itemsets is an active area in data mining that aims at searching interesting relationships between items in databases[11]. It can be used to address to a wide variety of problems such as discovering association rules, sequential patterns, correlations and much more. The proposed approach utilizes an efficient algorithm called MAFIA (Maximal Frequent Itemset Algorithm) which combines diverse old and new algorithmic ideas to form a practical algorithm. The proposed algorithm is employed for the extraction of association rules from the clustered dataset besides performing efficiently when the database consists of very long itemsets specifically.

Pseudo code for MAFIA :
MAFIA(C, MFI, Boolean IsHUT)
{
name HUT = C.head C.tail;
if HUT is in MFI
stop generation of children and return
Count all children, use PEP to trim the tail, and recorder by increasing support,
For each item i in C, trimmed_tail
{
IsHUT = whether i is the first item in the tail newNode = C I
MAFIA (newNode, MFI, IsHUT)
}
if (IsHUT and all extensions are frequent)

Stop search and go back up subtree
If (C is a leaf and C.head is not in MFI)
Add C.head to MFI
}

The cluster that contains data most relevant to heart attack is fed as input to MAFIA algorithm to mine the frequent patterns present in it. Then the significance weightage of each pattern is calculated using the approach described in the following subsection.

Significance Weightage Calculation

After mining the frequent patterns using MAFIA algorithm, the significance weightage of each pattern is calculated. It is calculated based on the weightage of each attribute present in the pattern and the frequency of each pattern.

$$S_{Wi} = \sum_{i=1}^{n} W_i f_i$$

Where *Wi* represents the weightage of each attribute and *fi* denotes the frequency of each rule. Subsequently the patterns having significant weightage greater than a predefined threshold are chosen to aid the prediction of heart attack

$SFP = \{x : Sw\ (x) = \Phi\}$

Where SFP represents significant frequent patterns and Φ represents the significant weightage. This SFP can be used in the design of heart attack prediction system.

6. CONCLUSION

In this paper the focus is on using different algorithms and combinations of several target attributes for effective heart attack prediction using data mining. Decision Tree has outperformed with 99.62% accuracy by using 15 attributes. Also the accuracy of the Decision Tree and Bayesian Classification further improves after applying genetic algorithm to reduce the actual data size to get the optimal subset of attribute sufficient for heart disease prediction.

Association classification technique apriori algorithm, was along with a new algorithm MAFIA was used .

Straight Apriori-based algorithms count all of the 2k subsets of each k-item set they discover, and thus do not scale for long item sets. They use “look a heads” to reduce the number of item sets to be counted. MAFIA is an improvement when the item sets in the database are very long.

REFERENCES

[1] P .K. Anooj, ―Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules‖; Journal of King Saud University – Computer and Information Sciences (2012) 24, 27–40.

[2] Nidhi Bhatla, Kiran Jyoti"An Analysis of Heart Disease Prediction using Different Data Mining Techniques".International Journal of Engineering Research & Technology
[3] Jyoti Soni Ujma Ansari Dipesh Sharma, Sunita Soni. "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction".
[4] Chaitrali S. Dangare Sulabha S. Apte, Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques" International Journal of Computer Applications (0975 – 888)
[5] Dane Bertram, Amy Voida, Saul Greenberg, Robert Walker, "Communication, Collaboration, and Bugs: The Social Nature of Issue Tracking in Small, Collocated Teams".
[6] M. Anbarasi, E. Anupriya, N.Ch.S.N.Iyengar, ―Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm‖; International Journal of Engineering Science and Technology, Vol. 2(10), 2010.
[7] Erickson, T., Smith, D., Kellogg, W., Laff, M., Richards, J and Bradner, E. Socially translucent conversations: Social proxies, persistent conversation, and the design of "Babble."Proc. ACM CHI (1999), 72–79.
[8] Hollan, J., Hutchins, E. and Kirsh, D. Distributed cognition: Toward a new foundation for human computer interaction research. ACM TOCHI, 7(2),(2000), 174–
[9] Shantakumar B.Patil, Y.S.Kumaraswamy, ―Intelligent and Effective Heart Attack Prediction System Using Data Mining and Artificial Neural Network‖; European Journal of Scientific Research, ISSN 1450-216X Vol.31 No.4, 2009.
[10] Statlog database: http://archive.ics.uci.edu/ml/machinelearning- databases/statlog/heart
[11] Shantakumar B.Patil,Dr.Y.S.Kumaraswamy "Extraction of Significant Patterns from Heart Disease Warehouses for Heart Attack Prediction" IJCSNS International Journal of Computer Science and Network Security, VOL.9 No.2, February 2009

INTENTIONAL BLANK

SHARING IS CARING: A DATA EXCHANGE FRAMEWORK FOR CO-LOCATED MOBILE APPS

Joseph Milazzo, Priyanka Bagade, Ayan Banerjee, and Sandeep K.S. Gupta

IMPACT LAB (https://impact.asu.edu/),School of Computing Information Decision Systems Engineering, Arizona State University, Tempe, Arizona
{joseph.milazzo,pbagade,abanerj3,sandeep.gupta}@asu.edu

ABSTRACT

Data sharing between mobile apps that are co-located in a smartphone may lead to synergistic benefits to the user especially in the health-care domain. Mobile apps that monitor and control user behavior can interact to engage the user into a healthy schedule over a long term. In the current app development paradigm, apps are being developed individually and agnostic of each other. To enable interaction between apps in secure manner, collaboration between developers is needed, which can be problematic on many levels. Current approaches to app integration require large code modifications to reap the benefits of shared data such as requiring developers to provide APIs or ensuring security through elaborate measures. In order to promote app interaction, this paper proposes a non-invasive secure interface for data sharing between mobile apps. A separate app, called the Health-Dev data manager, acts as a registry to allow apps to register database tables to be shared. Two health monitoring apps are developed to evaluate the sharing framework and different methods of data integration between apps. The health monitoring apps have shown non-invasive solutions can provide data sharing functionality without large code modifications and collaboration between developers.

KEYWORDS

Health apps, Data sharing, Mobile computing, Smart phone, Application Programming Interface, Data security, Data Management.

1. INTRODUCTION

Mobile computing [12] using mobile apps is becoming very prevalent. Mobile apps are being developed which are increasingly smart which use of context-sensing technologies. A user typically uses several related mobile apps on her smart phone. Data sharing between some of these multiple collocated mobile apps may make them more context aware with a relatively lower resource requirements. For example, let's consider different applications a) PETPeeves [1], which monitors the exercise patterns using accelerometer and Global Positioning System (GPS) data, and b) continuous electrocardiogram (ECG) monitoring application. If PETPeeves has access to

Sundarapandian et al. (Eds) : CCSEIT, DMDB, ICBB, MoWiN, AIAP - 2014
pp. 61–82, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4808

heart rate from the ECG app, it can compute calories burnt during exercise without interfacing with a heart sensor. In general, app integration has been found to be beneficial in increasing usage through methods such as experience modifiers and unlocking difficulty levels.

Unfortunately, in the current state, data sharing is not easily achievable due to data privacy issues [2]. If an app wishes to read data from another app, either one of two approaches is taken: the developers contact each other and share internal details of data storage including data read and write permissions; or the developer of the app whose data is being read must develop a custom Application Programming Interface (API) for other apps to access data. The API does not have the capability to restrict access to the data and hence may result in data leakage to unauthorized apps. A solution is to write custom APIs for each app attempting to access data, however, this places a burden on the developer to develop and maintain potentially large number of APIs.

In this paper, we propose a framework that enables app integration in a manner secure as well as non-invasive to the developer. The core of this framework is the Health-Dev [7] data manager app (HDDM) that maintains a registry of all available app databases. HDDM is considered as a trusted entity and any registered app opens up database access to the HDDM. Thus, instead of writing a custom API, the developer only needs to register tables of their database with the HDDM before another app can query details from the framework. Figures 1 and 2 illustrates how the framework improves data sharing. Without the framework (Figure 1), a custom API is developed for ECG monitoring app and thus PETPeeves can use the API to access the database. However, with the framework (Figure 2), ECG monitoring app first registers a table with the HDDM; PETPeeves can then query the HDDM for ECG monitoring app's database.

One interesting difference between two systems shown in Figures 1 and 2 is with the proposed framework external sensor communicates with the HDDM app rather than the ECG monitoring app. A benefit of this is that the sensor data may be shared among multiple apps in real-time. Thus, in addition to data sharing, the framework can interface sensors with apps. With the data manager handling the communication between the sensors and the apps, the possibility of another app making use of the same sensors' data exists and leads to shared functionality which is otherwise difficult to implement. This will be discussed further in Section III-A.

The proposed inter-app data sharing framework is non-invasive. Non-invasive in this context refers to minimizing required changes made in an app to achieve shared data. The framework allows multiple apps to become aware of user's activities and physiological data through shared data from other apps in only a few steps. To validate the framework, this paper uses a suite of apps, bHealthy [1], designed to share data to provide synergistic feedback and motivate the usage of multiple behavioral health monitoring apps.

An added advantage of this data sharing framework is that it can also be used as an interface to external sensors. The HDDM can communicate with multiple sensors, create databases sensed signals and provide simultaneous access to multiple apps.

Challenges: The principal challenges in data sharing between mobile apps are as follows:

1) Trade off between security and programming workload: In the current app development methodology, to share data in a privacy assured manner, app developers need to manually provide database access to individual apps. To reduce the burden of programming, app developers might make app databases globally accessible. However

that may give rise to serious security vulnerabilities [13]. The research question that we consider in this paper is, how to enable secure data sharing between apps without developer intervention?

2) synergistic feedback: When shared data is included into an app, the integration of data is vital to providing feedback.
How can shared data be integrated to produce synergistic feedback?

Contributions: Main contributions of the paper are summarized below:

1) Health-Dev data manager interface for secure data sharing between apps;
2) The HDDM service to interface sensor and app which can route sensor data to multiple apps.
3) Demonstration of synergistic benefits obtained due to secure sharing of data between apps;

2. EXISTING TECHNIQUES FOR APP INTEGRATION

This section discusses the current state of data sharing in mobile apps and the limitations these place on the advancement of data sharing in a plug-n-play system. This section first reviews the current state of inter-app communication across iOS and Android, the two most popular mobile operating systems. This section discusses the perceived limitations of the current state and compare related works against different granularity of changes needed.

A. Android

In Android data sharing between applications can happen if the data is stored in a structured format such as a data base table. As shown in Figure 3, to enable data access by any other app, the App 1 has to define a data base schema. This schema has to be agreed upon by the App 2 in order for it to read data from App 1. Hence at this step it requires collaboration between app developers to decide and design a common schema. On top of this schema App 1 has to define a content provider that provides access to the underlying data base. In recent Android OS every content provider is restricted to the App from which it is declared. In order to provide access to the content provider to App 2 the developer of App 1 has to set the access permissions. App 1 developer can either open up its content provider to all apps, which is often undesirable given the private nature of data, such as health data, or the App 1 can give access to only App 2. In such a case App 1 and App 2 developers need to collaborate and decide on a security mechanism. Based on the data base schema and the content provider security mechanisms the App 2 then implements a content resolver through which it can access data bases from the content provider of App 1. Once a content resolver is written in App 2, App 1 has to implement a broadcast messaging service to send notifications to App 2 whenever new data is ready. App 2 has to implement a broadcast receiver that will invoke the content resolver whenever new data from App 1 is available.

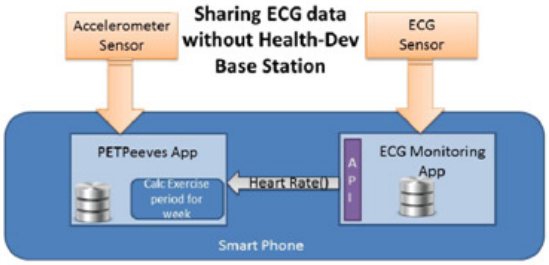


Fig. 1. System model of sharing meals table between apps without framework.

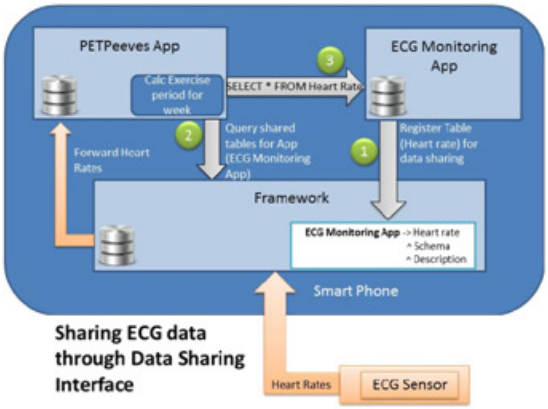


Fig. 2. System model of sharing meals table between apps with framework.

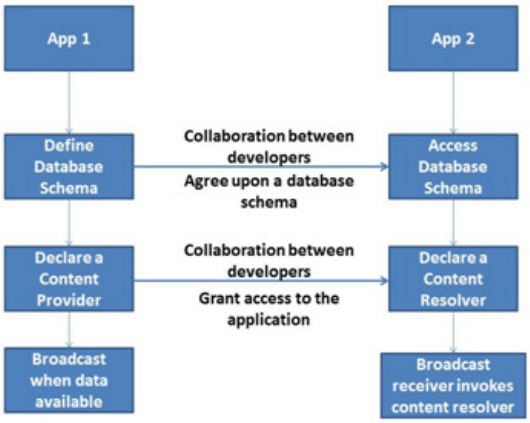


Fig. 3. Inter App interaction for Android platforms.

B. iOS

iOS runs apps in sandboxes which provide limited means of communications between apps directly on the device. Shared files and messaging systems are currently not present in iOS. However, iOS does provide the ability for apps to register URL Schemes. URL schemes are currently used by many apps to launch other apps and pass basic data through URL parameters,

but there is no structure to these URLs and no current standard for allowing callbacks to be passed in the URL; if the originating app wishes to receive some result based on the action.

The authors in [3] propose an Open Source library, Inter-App Communication, to expand on iOS lack of inter-app communication by providing callbacks to URL Schemes. This framework allows for callbacks to be registered based on the x-callback-url protocol specification. These callbacks play an important functionality for data sharing as they allow an app to develop an API for data sharing that can return results to the requesting app.

Table Comparison Of Different Interfaces

I. IN MOBILE SYSTEMS.

Interfaces	API	Structural Changes	Interoperable	Code Generated
Mobius	Required	Minimal	No	No
Simba	Required	Minimal	No	No
SOCAM	Required	Large	No	No
Health-Dev data man-Ager	Optional	None	Yes	Yes

C. Existing Works

Mobius is a middleware for interfacing complex data management schemes [4]. It provides programming abstractions of logical tables of data that spans devices and clouds. Applications using Mobius can asynchronously read and write these tables and receive notifications when tables change via continuous queries on the tables. The developer has a problem of complex handling of data management. There is no native solution from the platform, thus Mobius provides an interface to simplify data management. Another interface called Simba simplifies the complexities of synchronizing data with the cloud; hence reducing the developer's work [5]. The Service-Oriented Context-Aware Middleware (SOCAM) enables rapid prototyping of context-aware services in pervasive computing environments [6]. SOCAM provides a middleware of components to define context providers, interpreters, and the interaction between different components. Their middleware can allow data to be interpreted different between multiple apps; however each app must use the middleware to participate.

Table I provides an overview of the different interface approaches taken over the literature discussed. As the table describes, many solutions use an API to provide the interface and require an app to change structural design to adapt to the interface. None of the related works present a non-invasive interface which does not require extensive structural changes. In addition, such solutions are closed systems such that if an app does not use the interface, it is excluded from the benefits of the apps using the interface.

3. HEALTH-DEV DATA SHARING FRAMEWORK

The Health-Dev data manager is a background service running in a smartphone. It maintains two registries: a) registry for App databases, which maintains a log of the data base names and schemas that an App has made available for access, and b) a registry of permissions set for each shared data base. The methodology by which two apps can share data through Health-Dev is shown in Figure 4.

Two apps that want to communicate with each other should first register with the Health-Dev Data Manager service. The registration process ensures that the App and the Health-Dev service agrees upon a security protocol and a secure key. On creation of each database the App 1 broadcasts the name of the database and its schema to the Health-Dev service. The Health-Dev service stores the database schema and maintains the registry of databases. App 1 then implements content providers for each data base and uses the security primitives agreed upon with the Health-Dev to securely update the database in the Health-Dev data manager. Database updates are done using broadcasts between App 1 and the Health-Dev data manager.

Another App 2 which participates in the data sharing methodology also registers with Health-Dev and establishes a secure channel. App 2 can query the registry of Health-Dev and get a synopsis of the data bases that are available for access. App 2 can select the database that it wants to access and queries the Health-Dev registry to obtain the data base schema. The App 2 implements content resolvers for each database it wants to access. The App 2 then implements a broadcast receiver for each database and listens to database update broadcasts from the Health-Dev Data Manager.

A. Health-Dev Data Manager

The Health-Dev data manager is a background service in the smartphone. The primary functionality of Health-Dev data manager is to maintain a repository of all share-able database tables related to a registered app. Further, the data manager also handles the communication between sensors and application code, which is implemented using Health-Dev an automated code generator [7]. It exposes an Application Programming Interface (API), which allows a third-party application to register itself to receive sensor data, communicate with the sensor, add custom algorithms to run in the data manager, and lastly register custom callbacks such that the third-party app can be made aware of certain state changes in the data manager, such as loosing connectivity with the sensor.

Health-Dev data manager acts as a middleman between sensors and the smart phone; handling communication and signal processing is handled by the data manager; in addition, sensor data, whether processed or raw, can be forwarded to third-party apps; to provide custom business logic and integration of the data. Figure 5 provides an overview of how data is received and forwarded to a third-party application. The data manager first receives a START_SENSING message which prompts the data manager (already paired with sensor) to send a packet to sensor to begin sampling and sending data. As data is received, packets are parsed and raw data is passed to a pipeline of algorithms. These algorithms generally consist of different signal processing methods, but can also contain custom code registered by a third-party app. Once the algorithms are finished executing, the data manager checks to see if any apps have registered to receive data and if so, the processed data is broadcast to the app. The app is then free to use the data in whatever manner it wants. The data manager will continue this cycle until a STOP_SENSING message is received.

B. Inter-application Communication

Inter-application communication is handled through Broadcast and Broadcast Receivers in Android. Broadcast receivers are a core part of the Android OS and much of app development involves listening to various broadcasts from the Android OS to adapt to phone state, such as when the battery level updates. A Broadcast receiver is just a special type of listener for Broadcasts on the system. Broadcast receivers require a unique signature of a broadcast to be eligible to receive the broadcast. For example, the signature of a battery change event is Intent.ACTION_BATTERY_CHANGED. These signatures must be registered either during run-time or statically in the Android Manifest, a document which describes the app and permissions needed.

Broadcast receivers listen for broadcasts, which are structured messages which can be sent across the OS and received by an app. There are two types of broadcasts; normal or ordered. Normal broadcasts are completely asynchronous; all receivers are run in an undefined order, often at the same time. This is more efficient, but means that receivers cannot use the results or abort the broadcast. Ordered broadcasts are delivered to one receiver at a time. Each receiver executes in turn so it can propagate a result to the next receiver; or it can completely abort the broadcast so that it won't be passed to other receivers.

The data manager uses normal broadcast to broadcast data. Due to broadcast receivers needing to register for unique signatures of broadcasts, the data manager provides an API for registering and reading the different signatures. Health-Dev allows for sensor communication through direct API calls or through a broadcast API.

C. Database

The data manager also has the ability to store raw or processed data from the sensor in an internal database. This functionality allows for data to be buffered before broadcasting or simply to store if using the data manager as the app itself, instead of an API. In Android, content providers manage access to a structured set of data, such as a database. They encapsulate the data and provide mechanisms for defining data security, such as which apps can access the database.

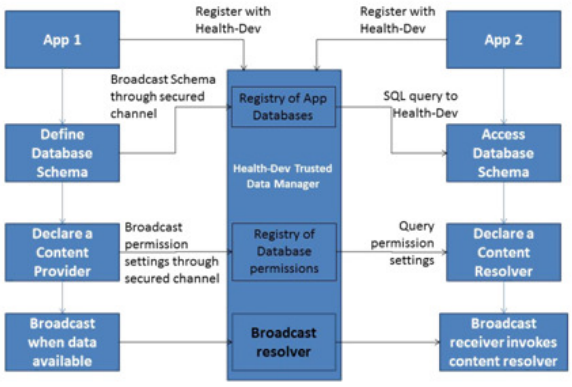


Fig. 4. Data sharing methodology for two apps using Health-Dev Data Manager.

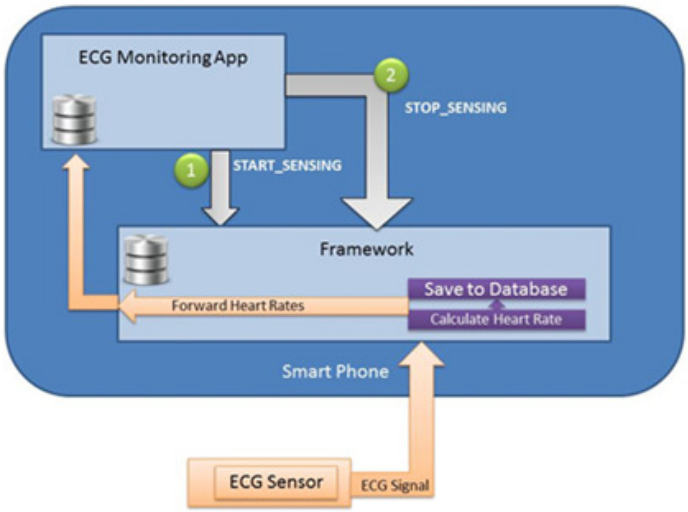


Fig. 5. Health-Dev data manager receiving data from external sensor and forwarding to app.

An application accesses the data from a content provider with a ContentResolver client object. This object has methods that call identically-named methods in the provider object. The ContentResolver object in the client application's process and the ContentProvider object in the application that owns the provider automatically handle inter-process communication.

In order for a client to access a ContentProvider, a content Uniform Resource Identifier (URI) is required and identifies data in a provider. Content URIs include the unique, name of the entire provider (authority) and a name that points to a table. An example of a content URI is content://user_dictionary/words. In this URI, the table name is "words" and "user dictionary" is the authority. The string "content://" is always present and identifies this as a content URI.

D. Limiting Visibility

Data sharing works by exposing portions of an app's data structure, in many cases tables in a database. For another app to query another content provider, a content URI must be known as well as the schema of that table and a description of each column. These are the core pieces of information an app must provide in order for an external query to be formed.

E. Participation

There is security concern with exposing database details in the open as well as restricting access after-the-fact. If the details were public knowledge, a change of content URI would be required as well as the complexity in changing database tables would increase. Instead, we propose using Health-Dev data manager as an interface to a registry, providing access to database details as well as allowing apps to enable or disable visibility of their details.

F. Implementation

The implementation of data sharing in data manager involves using Broadcast-based API. data manager creates a UUID (universally unique identifier) for each registered app. The app can use

an optional Schema Builder class, shown in Figure 6, provided by data manager API to pack the content URIs, schema, and descriptions into a custom object which is passed through a REGISTER_DATA_SHARING Broadcast. Upon receiving this broadcast, data manager generates the UUID and stores the passed information. At any time, the app can broadcast a DATA_SHARING_CHANGE which updates the access to the database details.

```
new BaseStationAPI.SchemaBuilder()
                  .createTable("meals")
                  .createColumn("calorie_intake")
                  .setType("integer")
                  .setRequired(true)
                  .setDefault(0)
                  .setDescription("Calorie intake for single meal")
                  .build();
```

Fig. 6. A Builder class provided to register database details with data manager.

A third-party app can read another app's database by first requesting a list of all app names registered. The app may then query data manager for a specific app's database details. The query is done by using the name of the app which data manager will translate into the unique identifier internally. Once database details are received, the third-party app can query the database as it normally would and use the results from there.

4. EXAMPLE APPLICATION SUITE: BHEALTHY

Healthy [1] is a suite of health monitoring apps which promotes data sharing for holistic health monitoring. The suite contains two applications which interact with mental and physical health. BrainHealth is the first app in the suite and monitors mental state to foster improved concentration, increase in mood, and reduction of stress through a technique known as Neurofeedback. BrainHealth integrates with PETPeeves, an app which promotes physical exercise through a virtual pet. The integration is detailed under PETPeeves section. The apps in bHealthy all use Health-Dev data manager for interacting with the external sensor.

A. BrainHealth

BrainHealth is a neurofeedback app which aids an user to learn how to permanently overcome behavioural problems, such as lack of focus, mood depression, and high stress. BrainHealth uses Electroencephalography (EEG) to extract the user's brain waves and interprets them as positive or negative for a chosen behavioral problem based on well-known protocols of Neurofeedback.

Neurofeedback has been found to be an effective method for encouraging healthy behavior. This app consists of three Neurofeedback Training activities: focus, mood change, and relaxation. Focus is aimed towards users who suffer from learning disabilities and need a boost in mental performance, motivation, and focus. Mood change is aimed towards users whom are not satisfied with their mood and want to achieve a more positive mood. Lastly, relaxation is aimed at any user who wants to learn how to relax in any situation.

System Architecture BrainHealth uses Emotiv EEG, a commercially available EEG headset which provides 14 channels. Due to Emotiv EEG using a proprietary communication medium,

direct communication between the EEG and Android phone is not feasible. Instead, the PC acts as a bridge to the phone; the EEG transmits raw data to the PC where the signal processing is performed on the data. The resulting output is transmitted to the smart phone app through Google Cloud Messaging. Google Cloud Messaging is a service in which messages are sent to Google's Cloud Platform. The messages are then delivered to the registered receiver via Wi-Fi or mobile networks.

Fig. 7. BrainHealth System Model.

Feedback Design BrainHealth's feedback loop is a particle system which manipulates particles spread out on the screen. When the user is performing well, the particles are attracted towards the center and combine with each other. However, when the user's performance degrades, the particles begin to split and spread towards the edges of the screen. Figure 8 shows how the particle system reacts to the user's brainwaves.

The feedback loop's particle system is manipulated by a single ratio derived from signal processing of EEG data against NFT protocols. The calculation, seen in Figure 7, first filters one second of data, then removing the DC offset and channels which are not relevant for the selected NFT activity. The data is then passed through a Hamming Window and each chunk has power spectral density (PSD) estimator ran on it. A ratio based on two bands is calculated from the PSD. The bands consist of one or more ranges of frequencies the user should excite or inhibit. To calculate the ratio, the PSD of the excitement band is taken over the PSD of the inhibit band.

The average ratio is stored in a database along with the NFT activity, total time, and date. This data can be used to track progress in BrainHealth and is shared to other apps such as PETPeeves, which will be discussed in the Section IV-B.

B. PETPeeves

PETPeeves is an app leveraging people's bond with a virtual pet. The app aims to encourage a user to increase or sustain physical exercise through bonding and caring for the virtual pet, shown in Figure 9. Several surveys, such as [8], have shown

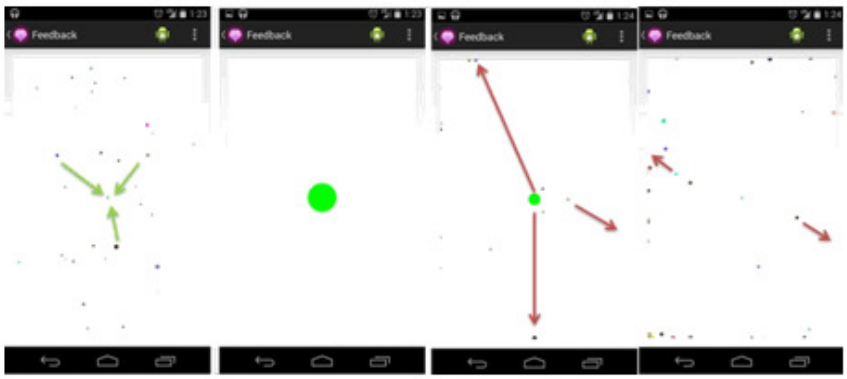


Fig. 8. Particles moving inward due to an increase in relaxation. Particles moving outward due to decrease in relaxation.

Table Ii.Pet's Experience Level Which Maps To A Specific Mood.

Levels	Mood
0 - 6	Unhealthy
7 - 16	Crummy
16 - 20	Neutral
21 - 28	Pumped
29+	Ecstatic

the effectiveness of virtual pets in encouraging positive mental state in children. PETPeeves manipulates the virtual pet's mood based on physical activity of the user.

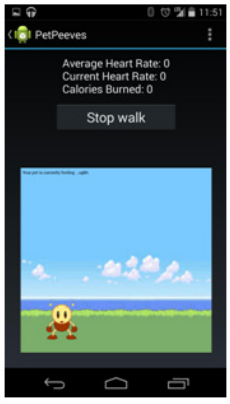


Fig. 9. Virutal Pet used in PETPeeves.

The goal is to motivate the user to exercise and keep their pet happy or to achieve the maximum level of happiness. If the user neglects the pet's happiness, the pet will becomes increasingly unhappy and the user will then need to work extra hard to achieve the previous level of happiness.

PETPeeves employs a leveling system for the pet such that as the user is exercising, experience points are earned. These experience points eventually cause the pet to level up. The amount of experience points increases depending upon the pet's level. For example, 17 experience is needed for a level 8 pet, but 20 experience is needed for a level 16 pet. A range of levels represent a specific pet mood, which can be seen in Table II.

The experience leveling formula is the same as a popular game, Minecraft created by [9]. The algorithm for calculating the amount of experience until the next level is:

During exercising, experience points are added and subtracted and eventually the pet will level up or down. Since experience

```
function experienceForNextLevel(currentLevel)
    if currentLevel ≥30 then
      return 62 + (currentLevel -30)* 7
    else if currentLevel ≥ 15) then
    return 17 + (currentLevel -15) *3

    else
    return17
    end if
    end function
```

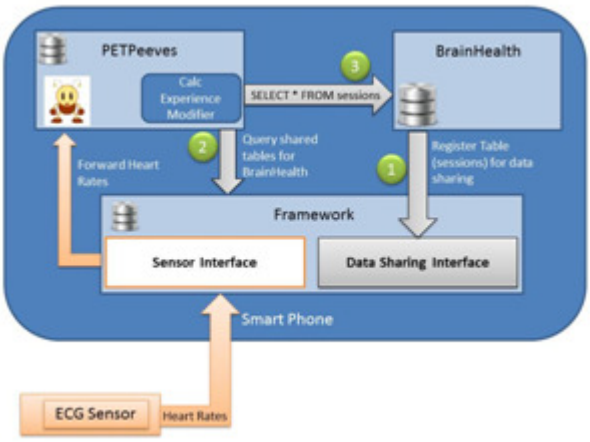


Fig. 10. PETPeeves and BrainHealth on the smart phone; PETPeeves is aggregating ECG data through sensor interface and querying BrainHealth's shared data through data sharing interface.

points are derived from physical activity, a health metric must be examined to determine if the user is exerting himself or not. The health metric used is heart rate extracted from ECG signals through an external sensor worn by the user.

5. APP INTEGRATION IMPLEMENTATION USING HEALTH-DEV

The HDDM is used to share data between PETPeeves and BrainHealth as shown in Figure 10. It shows the data sharing interface as well as the sensor interface. BrainHealth shares mental state data to PETPeeves. A positive mental health such as high focus level helps the user to get experience points in PETPeeves. Apart from the data sharing framework, the PETPeeves app uses the sensor interface to access the ECG sensor data.

Figure 11 depicts the application flow of the app. The user first selects a profile if already created. This switches to the exercising screen and loads the user and pet data. This information includes user's age and weight to calculate calories burned, and pet's experience levels. At this point, experience is deducted from the pet depending on the time since last app use and a modifier is given if the user has used BrainHealth within the past week. The user then presses the start button, which broadcasts a START_SENSING message to Health-Dev data manager and starts a timer. If there is no response from Health-Dev data manager, then the user is prompted to ensure the device is properly connected to and try again. If a response is received, a new Session is created. A Session is a period of time in which the user is engaged in either PETPeeves or BrainHealth. The Session consists of the feedback metrics used in the app and elapsed time of use. In the case of PETPeeves, the Session consists of date, elapsed time, pet's experience at the start and end of the session.

Since PETPeeves has asked to start sensing data from the ECG sensor, Health-Dev data manager is now broadcasting sensor updates to PETPeeves. In this case, the heart rate is calculated on the sensor and being forwarded to PETPeeves. On receiving of the sensor data message, the heart rate data is added to a rolling average and the current heart rate is updated. These pieces of data are then used to calculate the amount of experience to add to the pet, the mood is updated if needed, and the user interface is updated. After the user is satisfied with the progress made, the stop exercise button is pressed which stops the Session and broadcasts a STOP_SENSING message to Health-Dev data manager to stop sensing on the sensor. The session and pet data are updated in the database and the screen resets. The user is then free to close the app or start again.

A. Calculating Experience

At the start of a session, the pet's experience is modified based on the number of days since last used. If the app has not been used for more than a day, experience is subtracted from the pet such that the experience is

numberOfDaysSinceLastUse* experienceToNextLevel

PETPeeves calculates experience points to add/subtract on every heart rate update, which occurs once a second. During the HEART_RATE_UPDATE event, the new heart rate is added to a rolling average of 10 seconds as well as displaying the current heart rate to the user. The algorithm for calculating experience at each HEART_RATE_UPDATE event is shown below.

function calculateXPFromECG(averageHeartRate, baselineHeartRate)

delta ← averageHeartRate – baselineHeartRate

```
    if averageHeartRate ≤baselineHeartRate + 5.0
then
        xp ← -1

    else if delta ≤10.0 then xp ← -1

    else if delta ≤25.0 then xp ←1
    else

    xp ← 2 end if
    if moodModifier > 0 then

        bonusXP ← Math.round(Math.random() + moodModifier)

    else  bonuxXP ←0 end if

    return xp + bonusXP end function
```

B. Synergistic Feedback

PETPeeves takes advantage of shared data of BrainHealth through a bonus modifier. When BrainHealth is used in conjunction with PETPeeves, a modifier is granted which gives a small chance to earn an extra experience point during the session. If the user's average ratio in BrainHealth is larger than 0.5 (performs well in NFT), then there is a 10% chance to generate an additional experience per second, else there is a 5% chance is given.

6. SYSTEM VERIFICATION

Evaluation: The proposed framework consists of two interface modalities: 1) Sharing Data Interface and 2) Sensor Interface. These interfaces enable data sharing between apps and data forwarding between a sensor and multiple receiver apps. These interfaces heavily rely on the underlying Inter-Process Communication (IPC) mechanisms provided by Android. With the proposed framework with app suite, the data transmission latency is within accepted limit. This section focuses on, given a proposed framework, how latency scales with increased number of apps.

A. Setup

To evaluate the latency of the interfaces as the number of apps increase, three scenarios are considered as the usage models of the interfaces and served as the basis for the scalability model. These scenarios are derived from different usage scenarios of interfaces and are outlined as follows:

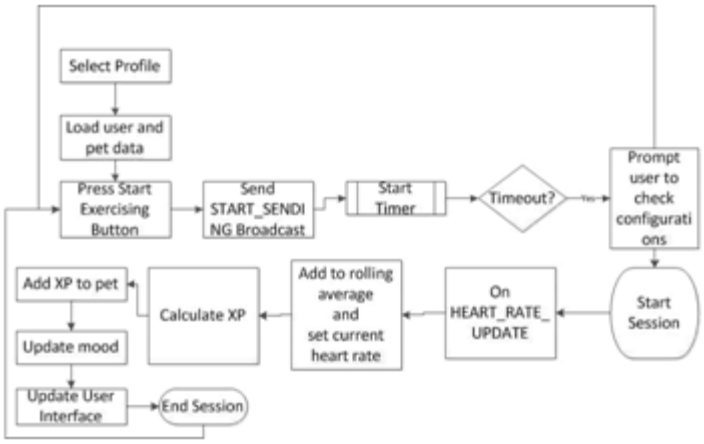


Fig. 11. PETPeeves application flow.

1) scenario 1: scenario 1 models two applications A_1 and A_2. Both applications receive X bytes of data every t seconds from the Sensor Interface. The IPC mechanism is synchronous such that a broadcast receiver order is followed and each app receives the broadcast in order from Sensor Interface.

2) scenario 2: scenario 2 models two applications A_1 and A_2. A_1 is receiving data through Broadcasts from Sensor Interface. A_2 is sharing data with A_1. There are 3 broadcasts and 1 Content Provider IPC call for data sharing. Like scenario 1, X bytes are received every t seconds from synchronous Sensor Interface broadcasts. The data sharing IPC calls are asynchronous, thus Android's scheduler handles the delivery and scheduling of the calls.

3) scenario 3: scenario 3 models apps A_1 and A_2 where A_2 is sharing data with A_1. One registration broadcast is made from A_2 and 2 broadcast and 1 Content Provider IPC call is made from A_1. This scenario only considers the Data Sharing Interface.

B. Model

Latency is dependent upon the scheduling of tasks (IPC calls) by the underlying scheduler of Android. Android is based on Linux 2.6 kernel and uses Complete Fairness Scheduler (CFS), which is an implementation of a well-studied algorithm, weighted fair queuing (WFQ). WFQ is a data packet scheduling technique allowing different scheduling priorities to statistically multiplex data flows. Each data flow is represented by a separate FIFO queue and an average data rate is calculated according to the equation below. In the equation, R is the total bandwidth in the system and w_i is the size of each queue.

$$Average\_Data\_Rate = \frac{\{R \times w_i\}}{\sum_i w_i} \qquad (1)$$

Using this equation for determining average bandwidth needed to keep all queues as balanced as possible, it can be applied to servicing a process' work. If each process has a queue with weight w_i, where w_i is the total size of work in queue (size of queue). The scheduler must balance the queues each time slice so that no process is neglected. We assume that Android's scheduler (CFS) operates in process detailed above.

The authors in [10] benchmark the IPC mechanisms provided by Android. In their results, latency increases significantly for IPC calls with a payload size over 4 KB. They note this is due to the initial kernel buffer being 4 KB and thus an allocation of temporary kernel buffer is required to transfer the payload. Bearing payload size in mind and considering the working of CFS, we have defined the latency model as:

$$L = \frac{\sum_{i=1}^{N} payload\_size_i + overhead}{R} \tag{2}$$

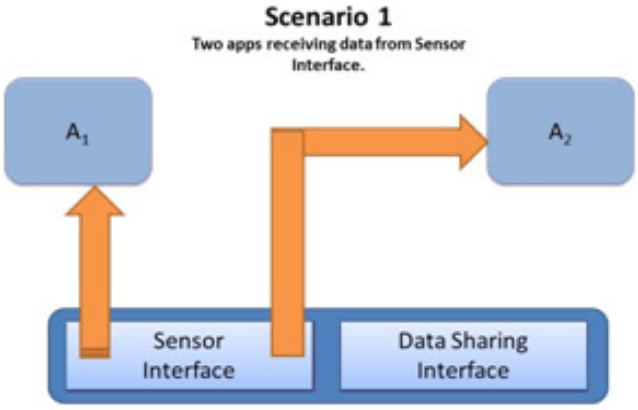


Fig. 12. Two apps receiving data from Sensor Interface.

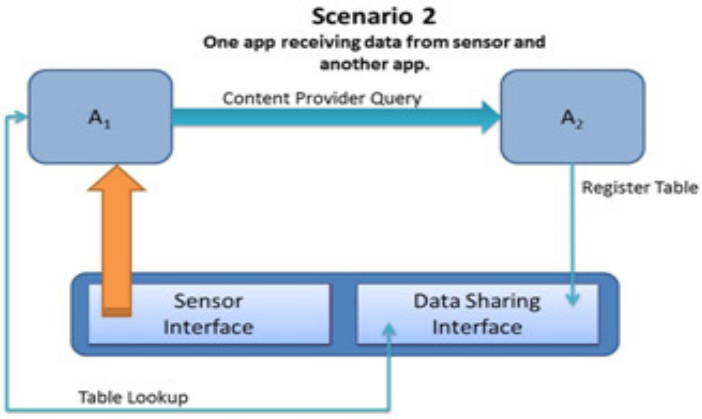


Fig. 13. One app receiving data from Sensor Interface and another app through Shared Data Interface.

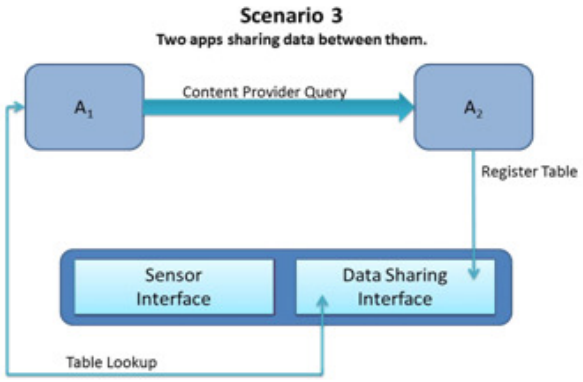


Fig. 14. Two apps sharing data through Shared Data Interface.

The total latency is the sum of all app's payload sizes (size of each queue) plus some overhead from the system divided by the total bandwidth the system provides. Overhead in smart phone can be created from events such as user interaction, incoming SMS, network connectivity, or phone calls. In Android, broadcasts that are not serviced within 1 second are prompted for killing. From this, it is assumed that if latency exceeds 10 second, the framework is not working correctly and is at the limit for scaling. Thus, L should be less than or equal to 10, L <= 10.

To validate the model, experiments were performed to collect latency data and test correctness of model proposed.

C. Experimental Setup

The experiments performed are to collect latency for IPC calls under a usage model of the framework. scenario 2 was chosen as the usage model and was broken into 3 experiments. For data sharing between apps, Content Provider IPC queries are used, while broadcasts are used for Sensor Interface data communication. For each experiment, the time for sending and receiving a broadcast is recorded. Broadcasts are sent once per second and payload sizes varies from 128 bytes to 256 KB.

Three total experiments are conducted and outlined below:

1) A_1 and A_2 communicate through data sharing framework. Content Provider IPC queries of a fixed size are sent once a second. There is no communication with the Sensor Interface.

2) Measurement of broadcast latency over varying payload size at 1 Hz between Sensor Interface and A_1. Payload sizes are: 128 bytes, 512 bytes, 1 KB, 4 KB, 16 KB, 256 KB. No data sharing takes place.

3) Data sharing and sensor interfacing. Content Provider calls are made at a fixed rate and payload while the sensor interface broadcasts different payload sizes

The experiments were performed on a Nexus 5 smart phone, the hardware specifications are shown in Table III.

TABLE III. NEXUS 5 HARDWARE SPECIFICATIONS.

CPU	2.3 GHz 4 Core Qualcomm Snapdragon 800 MSM8974
RAM	800 MHz 32-bit dual channel LP-DDR3 (12.8 GB/s)

D. Results

Figure 15 depicts the differences in broadcast latency with and without content provider queries. It can be seen that at a payload size of 4 KB, the latencies begin rising. This falls in line with [10], suggesting that latency increases over broadcast IPC once the initial kernel buffer is filled. The figure shows that content provider IPC calls does play a difference in latency of broadcasts, but the main variable is payload size. At 512 KB, the latency is greater for broadcast and content provider than with only broadcast. We believe this may be due to Garbage Collection, which introduces delays of 10-30 ms.

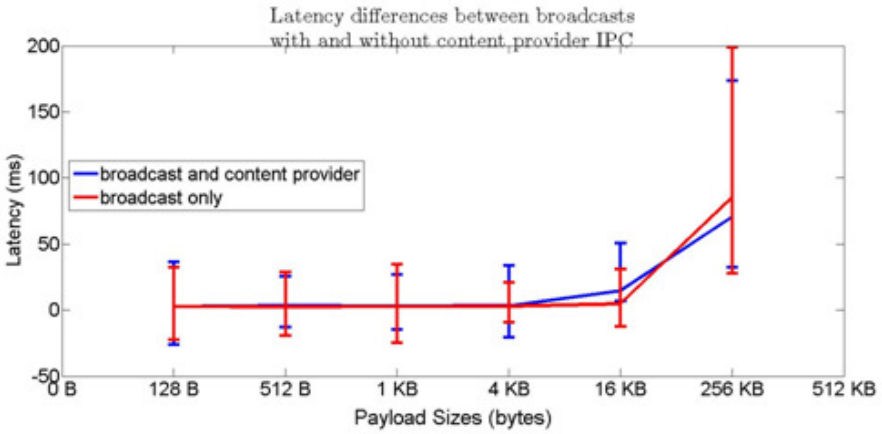


Fig. 15. Broadcast latency differences between inter-app communication with and without content provider IPC.

Content Provider IPC latency is compared with effect of broadcast and without in Figure 16. The latency without broadcasts produced from Data Sharing averages at 1 ms while without, the latency averages at 0.65 ms. Content providers are much faster than broadcasts, even with larger payload sizes (9 MB). However, the latency of a content provider is greatly affected by the payload of broadcast communication in the app. As the broadcast payload size increases, the latency also increases.

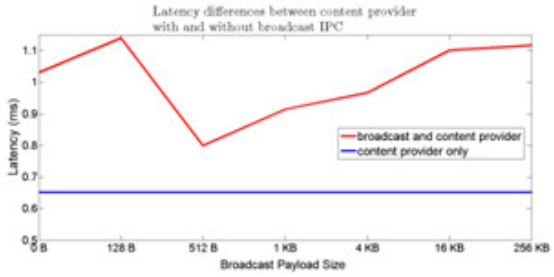


Fig. 16. Content Provider latency differences between inter-app communication with and without broadcast IPC.

7. DISCUSSIONS

A. Limitations

In this section, we discuss the concerns and feasibility of data sharing through the proposed framework. The limitations we discuss are data interoperability and multi-user environments. Data interoperability is the exchange of data between two parties in which both parties understand the data. A key issue between two separate systems communicating is how the data is formatted so that both systems can understand. In many systems today, XML or JSON are popular choices. These markups allow the data to describe itself including the structure.

In the proposed framework, data is not described in XML or JSON, however in a non-standardized way. To understand a query's data, the schema and description must first be retrieved from Data Sharing Interface and parsed. The parsing of the description is non-trivial due to having no standard and leaving the description to the registering app; thus a human reader is most likely necessary to fully understand the data format.

The reason for choosing a descriptor in the registry rather than wrapping the queried data in XML or JSON is to reduce work from the third party app, thus keeping the framework non-invasive. If the third party app were to wrap the result of a query in XML, they would be writing a basic API, which is one thing the framework aims to avoid. Instead, the user just needs to register the tables, schema, and write a short description of each field and their data is shared.

A limitation of the framework is multi-user support. An app's data is context-sensitive to the user account; there may be multiple user accounts per phone. This creates a problem to query for a specific user's data. There are two concerns: a) user information should be private to one app and b) there may be no similar fields between the two app's implementation of a user account. An example may be if one app uses OpenID, a framework for maintaining the same login to multiple apps, while another uses OAuth, a similar framework. Two different frameworks, no similarity between their user account data; how can these two apps be aware of which user is which?

This poses a severe limit on data sharing in a multi-user environment. A possible solution may be to store a unique identifier for a user in interface and have the app implement a translation method which will be called between queries, however this adds complexity from sharing data thus violating a core principle of the framework.

Cloud storage is another limitation of the framework. Many apps have begun storing their databases in the cloud with no local database. These apps are unable to share data in the current framework. While there may be solutions such as extending the interface to query the cloud's data, these come at the price of having an app implement an API. While the framework cannot operate with cloud-based databases, the solution presented does allow for offline availability.

B. Synergistic feedback

In PETPeeves, we have tried different techniques to integrate data shared from BrainHealth. These techniques were Compound Moods and Bonus Modifiers. Compound Moods was the addition of other moods based solely on the use of BrainHealth. An example would be if a user performed well in BrainHealth and was active in PETPeeves, his pet might be Focused and Fit. However no value was added to the experience of PETPeeves, just a different pet animation. Bonus Modifiers on the other hand change the experience of PETPeeves. For using BrainHealth recently, a user is granted a small chance to gain additional experience. The better the user performs in BrainHealth, the larger the bonus experience change is. As the pet levels, the pet becomes happier thus achieving the goal of the app.

During development of PETPeeves, two versions were developed. The first used Compound Moods while the other used Bonus Modifiers to integrate BrainHealth data into the app. we deployed the app on lab members and asked for feedback on the app as to whether they felt motivated while using to use BrainHealth. Compound Moods had negative reactions while we believe due to providing nothing for the user in return for using BrainHealth. However, Bonus Modifiers had good feedback especially when there was a progress bar so users could track their progress. We believe in order for synergistic feedback to occur from data sharing, that apps need to carefully consider how another app's data can provide something valuable to the user.

8. CONCLUSIONS

In this paper, we presented a secure non-invasive data sharing framework among apps to provide synergistic feedback for the user. The proposed solution requires minimal changes to an app to integrate and provides data sharing access to apps on a phone-wide scale. BrainHealth, a Neurofeedback training app shares neurofeedback data that PETPeeves uses through the data sharing interface. The data is used for possible bonus experience during PETPeeves use. This provides a synergistic feedback to the user.

The challenges faced were how to enable data sharing non-invasively and how to share only what is necessitated. The presented solution resolves both challenges. The data sharing framework uses a separate app as a registry for shared data access. The separated app allows for a number of apps to share or use shared data without making modifications larger than a few lines of code to their app. The separated app also ensures that as long as it is installed in the phone and at least one app has registered, then any app can query the shared data.

The second challenge is met through providing registration of specific tables to be shared, rather than whole databases. By also providing descriptions of each column in a table, some fields may be undocumented so as to protect possible information. While this challenge is met, there are possible complications that should be improved. This solution does not fully restrict access to

other tables of databases by not performing SELECT queries for the requesting app. By allowing the requesting app to make queries themselves, through carefully constructed queries, the app may gain knowledge of other tables in database.

The last challenge faced was how shared data can be integrated to produce synergistic feedback. This challenge is the combining of two apps data to enhance functionality or promote using two apps to supplement a user's lifestyle needs rather than one. This challenge is explored through bHealthy [1], the suite of health apps utilizing shared data from Neurofeedback to enhance physical exercise app. Compound moods and bonus modifiers were explored where bonus modifiers showed a positive response from lab members.

The interfaces proposed are non-invasive and automatically code generated. These aspects provide benefit such as faster development, reduction in human errors and effort, higher quality, and more control over how components, such as sensor and smart phone, interact and communicate with each other. The interfaces act as a separate app, which enable other apps to use shared data thus increasing the synergistic feedback.

Synergistic feedback through a collection of apps sharing data and adapting gives rise to apps that adjust and learn from other apps; new health apps which are aware of exercise or calorie intake recorded through other apps; and a system of non-fragmented apps.

An important component of any smartphone application is uploading data to a cloud server. In principle, data sharing between a smartphone application and a cloud service is similar to the problem of inter app data exchange. Although we have not explored solutions to the cloud integration problem in this work, an initial educated guess suggests that a solution through the usage of HDDM can be proposed. For example, the HDDM can allow registration of cloud services. In such a case, a cloud service can query the HDDM service in the phone regarding the availability of databases from apps. The secure communication of data between the HDDM and cloud service can be carried out using commonly used techniques such as https or digital signatures. In one of our previous work, we have proposed a solution that not only provides the cloud services access to databases obtained from physiological monitoring applications in the smartphone, but also encrypts the communication using the physiological signals. The physiological value based end to end security protocol (PEES) [11] shows the feasibility of non-invasive cloud integration through a trusted data manager such as HDDM.

Future Work: In this paper, different techniques of integration for shared data are proposed in hopes to produce synergistic feedback and promote a user to supplement their lifestyle with other health apps. These techniques were shown to lab members to gain feedback on how motivating each was, however there is no conclusive evidence that integration of shared data produce better results than using two apps separately. To extend this work, a study should be carried out to validate the hypothesis presented in this paper. Additionally, we will evaluate applicability of recently developed analytical techniques for developing smart mobile medical applications under dynamic context [14] application suites developed using this data sharing.

ACKNOWLEDGMENT

This research was funded in part by NSF grants CNS-1231590 and IIS-1116385. The work was done when Joseph Milazzo was in Arizona State University.

REFERENCES

[1] J. Milazzo, P. Bagade, A. Banerjee, and S. K. S. Gupta, "bHealthy: A physiological feedback-based mobile wellness application suite," in Proceedings of the conference on Wireless Health. ACM, 2013.

[2] S. K. S. Gupta, T. Mukherjee, and K. K. Venkatasubramanian, Body Area Networks: Safety, Security, and Sustainability. Cambridge University Press, 2013.

[3] A. C. Vivo, "Inter-app communication library," 2013. [Online]. Available: https://github.com/tapsandswipes/InterAppCommunication

[4] B.-G. Chun, C. Curino, R. Sears, A. Shraer, S. Madden, and R. Ramakrishnan, "Mobius: unified messaging and data serving for mobile apps," inProceedings of the 10th international conference on Mobile systems, applications, and services. ACM, 2012, pp. 141–154.

[5] N. Agrawal, A. Aranya, and C. Ungureanu, "Mobile data sync in a blink," in Presented as part of the 5th USENIX Workshop on Hot Topics in Storage and File Systems. San Jose, CA: USENIX, 2013. [Online]. Available: https://www.usenix.org/conference/hotstorage13/workshop-program/presentation/Agrawal

[6] T. Gu, H. K. Pung, and D. Q. Zhang, "A middleware for building context-aware mobile services," in Vehicular Technology Conference, 2004. VTC 2004-Spring. 2004 IEEE 59th, vol. 5. IEEE, 2004, pp. 2656–2660.

[7] A. Banerjee, S. Verma, P. Bagade, and S. K. S. Gupta, "Health-dev: Model based development pervasive health monitoring systems," in Wearable and Implantable Body Sensor Networks (BSN), 2012 Ninth International Conference on. IEEE, 2012, pp. 85–90.

[8] H. Kanoh, "Education for the net generation." [Online]. Available: http://www.childresearch.net/papers/digital/2008 01 01.html

[9] M. Persson, "Minecraft," 2014. [Online]. Available: https://minecraft.net/

[10] C.-K. Hsieh, H. Falaki, N. Ramanathan, H. Tangmunarunkit, and D. Estrin, "Performance evaluation of android ipc for continuous sensing applications," ACM SIGMOBILE Mobile Computing and Communications Review, vol. 16, no. 4, pp. 6–7, 2013.

[11] A. Banerjee, S. K. S. Gupta, and K. K. Venkatasubramanian, "PEES: physiology-based end-to-end security for mhealth," in Wireless Health, 2013.

[12] F. Adelstein, G. Richard, S. K. S. Gupta, and L. Schwiebert, "Fundamentals of Mobile and Pervasive Computing", McGraw Hill, 2004.

[13] A. Banerjee, K. Venkatasubramanian, T. Mukherjee, S. K. S. Gupta, "Ensuring safety, security, sustainability of cyber-physical systems," Proc. IEEE, 100(1), 2012.

[14] A. Banerjee and S. K. S. Gupta, "Analysis of Smart Mobile Applications under Dynamic Co̅nte̅xt," IEEE Trans. Mobile Computing, Aug. 2014 (Early access publication).

IMPLICIT CLIENT SIDE USER PROFILING FOR IMPROVING RELEVANCY OF SEARCH RESULTS

Saniya Zahoor[1] and Dr. Mangesh Bedekar[2]

[1] M.E., IIIrd Sem, Computer Engineering Department, MAEER'S MIT, Pune, India
saniyazmalik@yahoo.com
[2] Associate Professor, Computer Engineering Department, MAEER'S MIT, Pune, India
mangesh.bedekar@gmail.com

ABSTRACT

The Web is being a pool of knowledge, where any user visits hundreds of pages for various purposes but keeping track of its relevance for him is a tedious job. An average browser just provides you by the details of your browsing history but has no way to determine what importance the page holds for the user. In this paper we propose a method which aims to generate user profiles automatically depending on the various web pages a user browses over a period of time and the user's interaction with them. This automatically generated user profile assigns weights to web pages proportional to the user interactions on the webpage and thus indicates relevancy of web pages to the user based on these weights.

KEYWORDS

User profiling, web personalisation, implicit user behaviour modelling, client side analysis

1. INTRODUCTION

The Internet has made information available to humankind in a quick, easy, publicly accessible manner which is within reach of one and all. Today it's the first resource any user turns to when he needs any form of information. A multitude of web pages exist on any topic and the number grows with each passing day. The task of finding relevant information from the corresponding web pages is a tedious task. Search engines are available for the same but the problem of ever increasing data reduces the efficiency of the search results performed by a user. Even if the user manages to find a relevant page corresponding to his search query, retaining and remembering the webpage, or storing it efficiently for future references is another task not very well achieved. Identifying the relevance of a web page to a user thus becomes a very challenging task. Interest indicators abound but doing so implicitly is a challenge for any search system.

In this paper we propose a method which studies the user's behaviour on different web pages from search results. The method capture various user interest indicator parameters like, the mouse movement, mouse scroll, time spent on webpage and user actions like save, print and bookmark the webpage, which all users do unknowingly and analyse it to determine the relevancy of the webpage for the user. Once the relevancy is determined, ranking of the pages is done and stored

Sundarapandian et al. (Eds) : CCSEIT, DMDB, ICBB, MoWiN, AIAP - 2014
pp. 83–90, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4809

in a database which could be referred to, for any similar search query by the same user in the future.

Whenever a user gives a search query, the method scans the data stored in the database, with respect to the relevance of web pages, and gives the relevant search results from previous search queries, followed by other search results which would have come from the search engine normally.

We have proposed a technique of ranking web pages for a user according to its relevance to him by capturing six parameters all done on the client side. All this data is captured when the user accesses any webpage in the browser on his machine. This captured data is then analysed and relevance inferred. A result of testing our method proves that browsing for a user can be made more personalised and effective.

The main factors considered in this profiling are active time spent, pixels covered by the mouse pointer, vertical and horizontal scroll on a page and three others factors like save, bookmark and print. These six factors use statistical tools to allot weight to a given web page. We have made an extension to the formula already proposed in the paper [Teevan et. al., 2005]. We have also added factors like saving a web page, bookmarking a web page and printing a web page which are relevancy indicators too and explained the relevance factor in the calculations.

2. RELATED WORK

Many researchers have done work of allotting weights to the pages visited. This was primarily done on relating the number of pages a user visits and number of pages he finds relevant for a given search.

The method proposed by [1] ranks documents by summing over terms of interest the product of the term weight (wi) and the frequency with which that term appears in the document (tfi). When no relevance information is available, the term weight for term i is,

$$w_i = \log\left(\frac{N}{ni}\right)$$

Where N is the number of documents in the corpus, and ni is the number of documents in the corpus that contain the term i.

When relevance feedback information is available, two additional parameters are used to calculate the weight for each term. R is the number of documents for which relevance feedback has been provided. ri is the number of these documents that contain the term. The term weight in traditional feedback is modified as,

$$w_i = \log\left(\frac{(ri+0.5)(N-ni-R+ri+0.5)}{(ni-ri+0.5)(R-ri+0.5)}\right)$$

The methods proposed give relevancy of web pages to the user where both, feedback are un-available and whence relevance feedback is available.

In the earlier mentioned techniques, relevance of web pages to the user was gathered explicitly, by asking the user to rate the search results. Rating any web page explicitly after each search is a tedious task and is not in line with the usual behaviour of the user.

A lot of literature is available which speaks of how interest indicators can be gathered implicitly. [2] have compared web retrieval systems with explicit v/s implicit feedback. In their experimentation they show that the implicit feedback systems can indeed replace explicit feedback systems with little or no effect to the users search behaviour or task completion. [3] proposed to combine some well known interest indicators to get better results for relevancy of web pages, they also propose several more implicit interest indicators. [4] proposed a system to identify users' interest based on his behaviour in the browser he uses to access the Internet. They also go on to show that change in users interests can be handled by the system.

Considering the improvements suggested by the systems as mentioned above, we proposed modifications to the existing mechanism to infer interests of users on web pages implicitly by considering all the activities the users performs in the browser when he accesses a web page and to infer relevancy from it.

3. MODIFICATIONS PROPOSED

In our method we have proposed a relevance factor to determine if a web page is relevant or not, and if relevant how much. For this we sort pages in order of their relevant factors and only the top few pages (based on a threshold, to be taken from the user) are said to be relevant. The activity by the user on the webpage can be inferred by the active time spent by the user, mouse movement and scrolling behaviour on the webpage as indicated in [5] [6] [7] [8].

The relevance factor, Rf is, calculated by,

$$R_f = log\left(\frac{(\text{time}) \times (\text{movement})}{\text{scroll}}\right)$$

Time is the active time spent by a user on the web page. Active time is defined as the total time for which the user was on that page. The moment he switches to another tab / another window (the focus of the current open tab is lost) the timer stops and resumes only if and when the user comes back to the same tab. Time is calculated in seconds.

Movement is the number of pixels a user has moved over on the web page through his mouse pointer. Mouse movement hence suggests the importance of a web page to the user. Scroll is the total area of the web page scrolled by a user (horizontal scroll as well as vertical scroll),

$$\text{Scroll} = (V + v) \times (H + h)$$

V = height of the page
v = vertical scroll
H = width of the page
h = horizontal scroll

The other factors included are bookmarking [9], saving and printing [10] the web page. These factors are indicators that the web page contained relevant information for the user so much so that he might want to refer it in future and hence made a copy of a reference to it on his browser (machine). However these factors have different relevancies based on their inherent characteristics as described ahead.

Bookmarking – This is the most efficient way of keeping a reference to the web page for future. Bookmarking is able to handle all the changes which might happen on the web page as only a reference of web page is stored. This factor is considered better than save and print as the other two factors do not handle any changes which might have happened on a web page over a period

of time. Bookmark is also able to direct user to the web page in case the address changes over a period of time unlike saving and printing. Compared to the other two we allot a weight of '5 out of 10'.

Saving – Saving creates a copy of the web page on the user machine. Users prefer saving over printing as it saves paper and can be easily shared or edited. Though saving a page takes up memory space, it is still more preferable over printing. However saving cannot handle any changes made on that web page after it is saved. We allot a weight of '3 out of 10'.

Printing – This is another way how a user makes a permanent copy of the web page for future use. The user prints the web page from the browser if found important. However it is supposed to be a dead piece of information which doesn't incorporate any form of changes made to the web page after it gets printed. It also can't be trusted as there is no information about its source. Editing a printed document is difficult. Printing is also avoided by users as it consumes paper or a printer is not always available close by. We allot a weight of '2 out of 10'.

The moment the user saves and/or prints and/or bookmarks a web page the script running on the user's machine (client side), captures these actions and allots the relevant weight to the factor. These factor otherwise have a default weight of '1'. The importance of the accessed web page is determined by the user's action implicitly without ever asking him about the same explicitly.

We modify the relevance factor correspondingly to accommodate these user traits. The modified relevance factor is as mentioned below,

$$R_f = log\left(\frac{(time) \times (movement)}{scroll}\right) * save * print * bookmark$$

4. DESIGN CONSIDERATIONS AND IMPLEMENTATION DETAILS

The proposed method was implemented in the manner as explained below.

The implementation starts with installation of WAMP (Windows, Apache, MySQL, PHP) server on the client's machine and addition of a particular database with a table containing specific columns namely the URL, active time, mouse movement, scroll, save, print and bookmark. Here active time stores the value in seconds, mouse movement stores the value in pixels, scroll saves the value in pixels square and save print and bookmark stores the values of corresponding weights if the action occurs or it stores zero.

Once the server has been installed with the required database and table, every time the user logs into a browser he needs to switch on the server after which the method starts getting implemented as the data starts getting stored in the database after which its analysis starts.

Greasemonkey is a Mozilla Firefox extension that allows users to install scripts that make on-the-fly changes to web page content after or before the page is loaded in the browser (also known as augmented browsing).The changes made to the web pages are executed every time the page is viewed, making them effectively permanent for the user running the script. Greasemonkey can be used for customizing page appearance, adding new functions to web pages (for example, embedding price comparisons within shopping sites), fixing rendering bugs, combining data from multiple web pages, and numerous other purposes. With the help of greasemonkey we installed out script in all the web pages a user visits through Mozilla Firefox so that we could capture all the factors needed for the methods implementation.

Mozilla Firefox Browser is free and open source, its features include tabbed browsing, spell checking, incremental find, live bookmarking, smart bookmarks, a download manager, private browsing, Functions can be added through extensions, created by third-party developers, of which there is a wide selection, a feature that has attracted many of Firefox's users. One of the main reasons why Mozilla Firefox was used was because of its unique add-on Greasemonkey.

JavaScript (JS) is an interpreted computer programming language. It was implemented as part of web browsers so that client-side scripts could interact with the user, control the browser, communicate asynchronously, and alter the document content that was displayed. Javascript can be imbedded in the html page and have pre-defined function which help in capturing various data from the web page. Javascript along with tools of PHP and AJAX capturing the data needed for the method and directs it to the database where it gets stored.

WAMP is a form of mini-server that can run on almost any Windows Operating System. WAMP includes Apache 2, PHP 5 and MySQL (phpMyAdmin and SQLitemanager are installed to manage your databases) preinstalled. The WAMP server needs to be hosted every time the script is run. The javascript directs the data captured to the PHP page which stores the data in the corresponding table of the database in the WAMP server.

5. SCRIPT DESCRIPTIONS

The important script written in greasemonkey ensured that the moment the page loads the script gets implemented in all the web page. Hence this assures that the script runs in all the pages a user visits.

The scripts are modularised and handle user events. As soon as the webpage is completely loaded, a script starts calculating active time. Another script handles mouse movements. It returns the total mouse movement on the page. A third script handles any scrolls on the page. It handles both vertical and horizontal scroll. The counter decrements if there is any backward or upward scroll, the counter decrements. This gives the total scroll on the web page.

Three more user events, namely save, print and bookmark have been defined whenever the user saves prints or bookmarks a page. This script handles what all actions are performed by the user on the webpage.

The method as explained above was implemented on a Firefox browser using tools like html, PHP, JavaScript, JQuery, AJAX on Greasemonkey (an add-on available on Mozilla Firefox which allows to customize the way a web page displays or behaves, by using small bits of JavaScript) and hosted by WAMP server. The database was stored and retrieved from MYSQL.

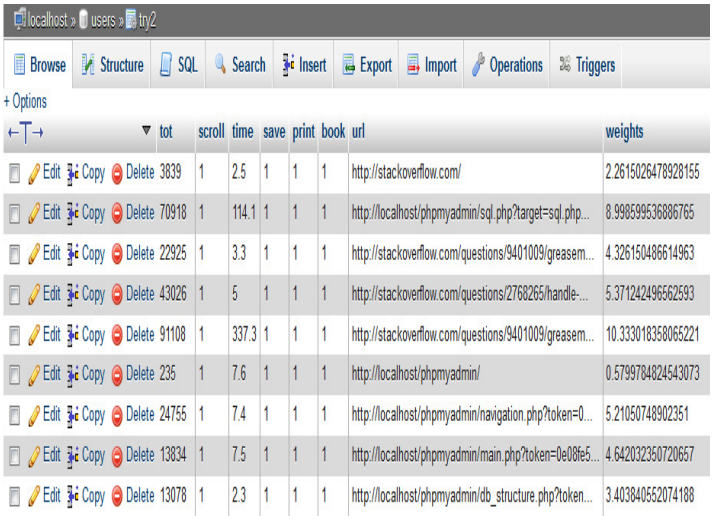
Scripts are written to calculate the active time a user spends on a given page. The mouse movement of a user was calculated by another script. Mouse movement was calculated using JQuery where the number of pixels (height * width) the mouse hovers on was calculated. The moment the document gets loaded a function keeps track for any sort of mouse movement.

Page Scroll was also calculated the same way using JQuery. Scroll refers to horizontal and vertical scrolling of the web page. The script also takes care that if the user scrolls downward and then upward the script nullifies the value of downward scroll with the upward scroll. This takes care of the factor that the content at some portion of the page wasn’t useful for the user because of which he came back to the initial position. The same goes for horizontal, forward and backward scroll.

Another script took care of keyboard inputs for page down, page up, end and home buttons. The scroll action from these buttons is also calculated. Keyboard Shortcuts for bookmarking, saving and printing were captured the same way where a function is triggered as soon as a keyboard stroke is recognized.

Once all these user initiated events are captured, this data along with the URL of the web page is stored into the database. The relevance factor script is then invoked using Gresemonkey which monitors the users' behaviour on the web page. In this way all the necessary data values required for the method are captured and stored.

Since the databases are stored on WAMP server, as shown in figure-1, which is hosted on a client's machine, the database is accessible only to the client who may password protect it, which can prevent any form of invasion of privacy.


localhost » users » try2

Browse | Structure | SQL | Search | Insert | Export | Import | Operations | Triggers

+ Options

	tot	scroll	time	save	print	book	url	weights
Edit Copy Delete	3839	1	2.5	1	1	1	http://stackoverflow.com/	2.2615026478928155
Edit Copy Delete	70918	1	114.1	1	1	1	http://localhost/phpmyadmin/sql.php?target=sql.php...	8.998599536886765
Edit Copy Delete	22925	1	3.3	1	1	1	http://stackoverflow.com/questions/9401009/greasem...	4.326150486614963
Edit Copy Delete	43026	1	5	1	1	1	http://stackoverflow.com/questions/2768265/handle-...	5.37124249662593
Edit Copy Delete	91108	1	337.3	1	1	1	http://stackoverflow.com/questions/9401009/greasem...	10.333018358065221
Edit Copy Delete	235	1	7.6	1	1	1	http://localhost/phpmyadmin/	0.579978482454307 3
Edit Copy Delete	24755	1	7.4	1	1	1	http://localhost/phpmyadmin/navigation.php?token=0...	5.21050748902351
Edit Copy Delete	13834	1	7.5	1	1	1	http://localhost/phpmyadmin/main.php?token=0e08fe5...	4.642032350720657
Edit Copy Delete	13078	1	2.3	1	1	1	http://localhost/phpmyadmin/db_structure.php?token...	3.403840552074188

Figure 1 – Screen shot of the table storing the data about user's activity on a web page.

6. DISCUSSION

Whenever a user gets the results for any search query, we can segregate that search query appearing on a page into different segments by using Greasemonkey scripts. These search result segments of a page are thus available to us, which can then be evaluated separately. Depending upon the amount of mouse movement which appears on a segment for the session of the query and detected amount of the mouse click on that segment, we can infer if the given search result was relevant to the user or not.

If the user scrolled till, say result number 5, out of the 10 results appearing on that web page, we can conclude that only the fifth or the top 5 results were relevant to the user. Furthermore, depending on how many segments had mouse hovering or clicks in it, we can conclude which all results among these were relevant to the user and how much. Using our method we can further re-rank search results on a search result page.

7. THE EVALUATION FRAMEWORK

The moment a user visits any web page the script captures it's URL along with the required six factors namely - time, scroll, movement, save, print and bookmark and stores it in the database as

soon as the web page is unloaded. The formula proposed above for calculating the relevance factors would be applied to different pages and pages will be sorted according to their relevance factor. This entry is stored for each session. Moreover these entries can be analysed for each day, week, month etc. and can help in depicting the change in a user's browsing pattern and interests over a period of time. Pages which have not been accessed over a long span of time (based on the threshold value) would be automatically truncated in order to increase the efficiency of the method and reduce the sample space evaluation.

8. RESULTS OBTAINED

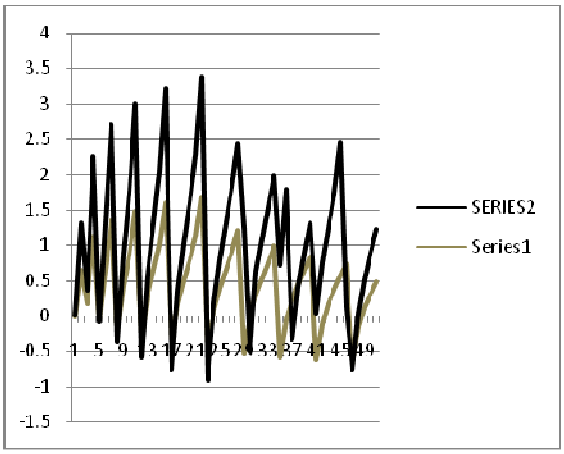


Figure 2 Comparison of performance measures

As depicted in Figure 2, the graph was obtained when we plotted the model initially proposed in [1] as series 1 and our model after the modification as series 2. We observed that the weight (on the y axis) obtained after our modification were lying in a larger range as compared to the model initially proposed.

The weights in model proposed in [1] was in the range of [-0.6, 1.75] whereas the weights obtained from our model after the modification were in the range of [-0.85, 3.4]. Hence, we could conclude that the weights after the modification were more discrete and well-spaced. Since the values of the weights were not very close to each other as earlier, the pages could be ranked in a better way.

9. INFERENCES

The model proposed in [1] did not handle user actions on the web page. However, as a human trait the actions of the user in the browser on the web page reveal the relevance of the web page to him. These user actions are implicit in nature and happen naturally which do not obstruct his flow of actions, if otherwise asked explicitly [11]. The proposed extension handles common user types like – a user who would bookmark a web page for future reference more often than saving or printing, a user who would save the web page for further reference, a user who would print the web page for future reference and a combination thereof.

10. FUTURE WORK

Since the framework can rank pages according to the user's relevance, this method will be used in giving relevant search results to the user. User profile thus generated will be used to give relevant

search to a user combined with the ranking returned by a web search engine. For each search term the pages browsed by the user are recorded and ranked according to his profile. Once a concrete database gets created over a period of time, as soon as the user searches any term he will get a list of pages visited by him for similar search done earlier which would be ranked according to their relevance, followed by the search results of the default search engine. In case the search is entirely new to his profile he would just get the search results of the default search engines, the framework will learn this new search query.

The relevance of the corresponding webpage can be further increased if the user copies text from the web page on to the clipboard, it indicates the relevancy of the contents of the webpage and hence the relevancy of the webpage itself as suggested in [12].

REFERENCES

[1] [Teevan et. al., 2005] Jaime Teevan, Susan Dumais, and Eric Horvitz, Personalizing search via automated analysis of interests and activities, In Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '05). ACM, New York, NY, USA, Pp. 449-456.

[2][Ruthven, 2002] White, R., Ruthven, I. and Jose, J.M., The use of implicit evidence for relevance feedback in web retrieval, Proceedings of the Twenty-Fourth European Colloquium on Information Retrieval Research (ECIR '02). Lecture Notes in Computer Science. Glasgow. 2002, Pp.93-109.

[3] [Shapira, 2006] B. Shapira, M. Taieb-Maimon, and A. Moskowitz. Study of the usefulness of known and new implicit indicators and their optimal combination for accurate inference of users interests. Proc. SAC '06, Pp.1118–1119.

[4] [Li, 2008] Fang Li, Yihong Li, Yanchen Wu, Kai Zhou, Feng Li, Xinguang Wang. Discovery of a User Interests on the Internet, In Proceedings of the IEEE/WIC/ACM, International Conference on Web Intelligence and Intelligent Agent Technology [C], 2008, Pp.359-362.

[5] [Kellar, 2004] Kellar, M., Watters, C., Duffy, J., & Shepherd, M. (2004). Effect of task on time spent reading as an implicit measure of interest Proceedings of the 67th Annual Meeting of the American Society for Information Science, 41, Pp.168–175.

[6] [Nagy, 2009] Istvan K. Nagy and Csaba Gaspar-Papanek, User Behaviour Analysis Based on Time Spent on Web Pages, Web Mining Applications in E-commerce and E-Services, Studies in Computational Intelligence, 2009, Volume 172 / 2009, Springer, Pp. 117-136.

[7] [Roman, 2011] Pablo Enrique Roman Asenjo, Web User Behavior Analysis, Doctoral Thesis, March 2011.

[8] [Nyman, 2013] Mathias Nyman, Navigation Behavior Analysis and User Profiling Based on Automatically Collected Website Data, Master's Thesis, School of Science, Aalto University, Finland, February 1, 2013.

[9] [Claypool, 2001] Mark Claypool, Phong Le, Makoto Wased, David Brown, "Implicit interest indicators", Proceedings of the 6th international conference on Intelligent user interfaces, January 14-17, 2001, Santa Fe, New Mexico, USA. Pp. 33-40.

[10] [Kim, 2005] Kim, H. and Chan, P. K. Implicit indicator for interesting web pages, International Conference on Web Information Systems and Technologies, Miami, 2005, Pp.270-277.

[11] [Reber, 1989] Implicit Learning and Tacit Knowledge Journal of Experimental Psychology: Arthur S. Reber (1989) General 1989, Vol. 118, No. 3, Pp. 219-235.

[12] [Holub, 2010] Michal Holub, Maria Bielikova, Estimation of user interest in visited web page, Proceedings of the 19th international conference on World Wide Web, WWW '10, April 26-30, 2010, Raleigh, North Carolina, USA, Pp. 1111-1112.

AUTHOR *INDEX*