PROCEEDINGS OF THE 23rd ANNUAL TECHNICAL CONFERENCE OF IET SRI LANKA NETWORK 2016

ISSN: 2345-9557

PROCEEDINGS OF THE

23rd ANNUAL TECHNICAL CONFERENCE OF IET SRI LANKA NETWORK

2016

Saturday 27th August 2016
Hotel Galadari Colombo, Sri Lanka

Organized by

Disclaimer

The views expressed in the papers published in these proceedings are those of the authors. The statements and opinions stated either in the papers published herein or at the conference and discussions do not necessarily represent the views of the IET Sri Lanka Network.

Proceedings of the IET Sri Lanka Network 2016

A publication of the Annual Conference of IET Sri Lanka Network

ISSN: 2345-9557

IET Sri Lanka Committee 2015 / 2016

1. Eng. Dammika Nanayakkara - Chairman
2. Eng. Rukmal Jayasinghe
3. Eng. Harshana Ranmuthugala
4. Eng. N.A.G Senanayake
5. Eng. Ravi Rupasinghe
6. Eng. Y.M Samarasinghe
7. Eng. Tharaka Mohotti
8. Eng. Rohan Seneviratne
9. Eng. K.N. Gunawardena
10. Eng. Prof. K.K.Y.W. Perera
11. Eng. Lakmal SenanayakeIPRA
12. Eng.W.D.de Alwis
13. Eng. D.R Pullaperuma
14. Eng. Prof. J.R Lucas
15. Eng. D.R.N. Jayasuriya
16. Eng. Dr. Lalith Liyanage
17. Eng. P. M. Wickramasinghe
18. Eng. Mahesh Thilakasiri
19. Eng. L. N. S. Perera
20. Eng. Anuradha Jayakody
21. Eng. Dayan Gunawardana
22. Eng. Dasitha Liyanage

Conference Committee 2016

Eng. Dr. Lalith Liyanage - Chairman
Eng. Anuradha Jayakody
Eng. Ravi Rupasinghe
Eng. Dayan Gunawardena,
Eng. Dasitha Liyanage
Eng. Naleen Jayasooriya

Editorial Board 2016

Eng Anuradha Jayakody – Chief Editor
Eng. Dr. Lalith Liyanage
Eng. Dayan Gunawardena

Panel of Referees & Judges

Prof. Rohan Lucas - Chairman
Mr D R Pullaperuma
Dr. Chulanthe Kulasekara

IET SRI LANKA NETWORK
23rd ANNUAL TECHNICAL CONFERENCE
SATURDAY 27th August 2016
HOTEL GALADARI, COLOMBO, SRI LANKA

Collective inspiration

Program

8:00 am to 8:45 am	Registration
8:45 am to 8:55 am	Welcome Address Eng. Dhammika Nanyakkara,Chairman IET Sri Lanka Network

Session A- Chair – Eng Tharaka Mohotty

8:55 am to 9:10 am	**A1: Androsafe: Online malware analysis with statistic and dynamic methods** Krishnadeva Kesavan , Chethana Liyanapathirana , Gayani Sadamali , S.A.W.S Sampath, Y.M. Sureni Koshila , Chamod Premarathne , Prabath Lakmal Rupasinghe, Chamira Nawarathna
9:10 am to 9:25 am	**A2: Database Optimization Model to Reduce Application Delay Relates to Mobile Devices** J.A.D.C. Anuradha Jayakody, D.S.A.Kandawala, K.D.Dinusha Chathurangi, N.T Mapa, T.S Disanayaka
9:25 am to 9:40 am	**A3: A Sentiment Analysis and Classification Algorithm utilizing an Independent Term Matching Scheme sensitive to Negations and Word Count Patterns** Dr. Asoka Korale, Chanuka Perera, Eranda Adikari, Nadeesha Ekanayake
9:40 am to 9:55 am	**A4: A Light Weight Provenance Aware Trust Negotiation Algorithm for Smart Objects in IoT**

	J.A.D.C. Anuradha Jayakody, Prabath Lakmal Rupasinghe, N.T Mapa, T.S Disanayaka, D.S.A.Kandawala, K.D.Dinusha Chathurangi, Krishnadeva Kesavan
9:55 am to 10:10 am	**A5: The Simulation – Based Solution to Detection by Collectively Decomposing Factors of IMU and Image Process Data in the Indoor Environment** A.D.C.Anuradha Jayakody,R.G.P Dulshan, D.C Meegahawatta, H.M.K.S.B Herath, L.P.A.D Pathirathne
10:10 am to 10:25 am	**A6: Predictive Analytics with online data for WSO2 Machine Learner with the support of Ensemble method** Prabath Lakmal Rupasinghe, Heshani Herath, Ishani Pathinayake, Ashani Diaz, Indujayani Karthigesu, Krishnadeva Kesavan , Chethana Liyanapathirana, Sripa Vimuthi
10:25 am to 10:40 am	**A7: NExT-Plugin for MetaTrader 5 based on Recurrent Neural Networks** H.A.C.J Hettiarachchi, L.I Hettiarachchi, M.M.N.M.Mullegama, R.K.T.D Ranaweera, Prabath Lakmal.Rupasinghe, D. Kasthurirathne
10:40 am to 10:55 am	**A8: Mitigation of the Evil Influence of Internet Use on Teenagers in Sri Lanka** T.G. Premarathne, Yasas Mallawaarchchi

Session B - Chair –Eng. Rukmal Jayasinghe

	B1: A Comparative Analysis on Technologies Available for Localization in Indoor Navigation
11:15 am to 11:30 am	J.A.D.C.Anuradha Jayakody, Shashika Lokuliyana, K.Lakshmi Lashini, S.A.Rangala.
11:30 am to 12:00 noon	**B2: ClusterMal: Automated Malware Analysis with**

	clustering, anomaly detection and classification of existing and new behavioral analysis Krishnadeva Kesavan, Sripa Vimukthi Bannakkotuwa, V.V.YY. Wickramanayake, M.P.D.H De Silva, J.M.D. Fernando, K.K.K.K. Sampath, Prabath Lakmal Rupasinghe
12:00 noon to 12:15 pm	**B3: Reduction of power losses in LV distribution network by improving the distribution line connection.** Eng. K.Lingaruban, Prof.M.A.R.M.Fernando, Dr. A.Atputharajah, Eng. D.K.P.U.Gunathilake
12:15 pm to 12:30 pm	**B4: Engineer adoption of new technology to boost social development: A perspective on digitalization of TV broadcasting in Sri Lanka.** Eng.Kumudu wasantha Muthuthanthrige
12:30 pm to 12:45 pm	**B5: Power Analysis Based Side Channel Attacks** Hasindu Gamaarachchi, Harsha Ganegoda, Roshan Ragel
12:45 pm to 1:00 pm	**B6: Fuzzy Logic based Motion Controller for Underwater Remotely Operated Vehicle** P.M. Saranguhewa, S.D.S. Lakmal, R.M.T. Lakmali, N.D. Jayasundere, S.H.K.K. Gunawickrama
1:00 pm	**Announcement of the Best Paper - Panel of Judges** **Award of Author Certificates** **Vote of Thanks** Eng. Dr. Lalith Liyanage – Conference Chairman

Message from the Chairman – IET Sri Lanka Network 2015/2016

The Annual Technical Conference is organized by the IET Sri Lanka Network for the 23rd consecutive time. The Institution of Engineering and Technology - IET, (formerly known as the Institution of Electrical Engineers - IEE), is the largest professional engineering society in Europe which has more than 160,000 members in 127 countries and offices in Europe, North America, and Asia-Pacific. Having established as way back as 1871, IET is one of the oldest engineering institutions in the world. The IET provides a global professional network to facilitate the exchange of knowledge and ideas and promotes the positive role of Science, Engineering, and Technology to engineer a better world.

Holding a successful technical conference year after year constitutes a hallmark of excellence for a professional society. IET Sri Lanka Network organized a very successful international conference on the Internet of Things (IoT) last month, first of its kind in Sri Lanka and the conference was highly commended by more than 250 industry professionals attended the event. While many factors contribute to the success of such a conference, a key factor is careful planning and organizing of events in advance. The purpose of the annual technical conference is to honour outstanding technical paper presentations and publication in the fields of Electrical, Electronics, Telecommunications, Civil, Mechanical, Automobile, Marine, Aeronautical, Manufacturing and Information Technology research and innovations and practice. This conference is open to professionals in industry and academia.

I would like to express our deepest appreciation to the panelists Professor Rohan Lucas, Mr. D R Pullaperuma and Dr. Chulantha Kulasekara. We are most grateful to the reviewers Prof Rohan Munasinghe, Dr Indika Perera and Dr. D. A. Meedeniya from the University of Moratuwa, Dr. Nihal Kodikara from the University of Colombo and Dr. Pradeep Abeygunawardhana from Sri Lanka Institute of Information Technology, who have so diligently supported the paper review process. I would like to specially thank the Committee Chairman, Dr. Lalith Liyanage and his team, the sessions Chairmen, technical paper presenters and the entire organizing committee. The success of the Annual conference would not have been possible without their tremendous volunteer effort.

I wish this year's Conference and presenters all the very best and look forward to seeing you at the Next Annual Conference.

Eng. Dammika Nanayakkara, *BSc.Eng(Hon), MEng, CEng,(UK), MIE(SL), MIEEE, FIET, IntPE(UK)*

Hon Chairman
IET Sri Lanka Network.

Message from the Chairman – Conference Committee

On behalf of the Organizing Committee, I warmly welcome you to the 23rd Annual Conference of the Institution of Engineering and Technology (IET), Sri Lanka Network. The objective of holding this conference is to encourage professionals in the fields of Electrical, Electronic, Telecommunication, Civil, Mechanical, Automobile, Marine, Aeronautical, Manufacturing and Information Engineering in Sri Lanka to meet with each other, exchange ideas and share information on research, development, and innovations taking place in the country.

There are Fourteen (14) high standard technical papers on various engineering aspects which have been presented and accepted by the referees are published in this edition of the Conference Proceedings of the IET Sri Lanka Network 2016.

I wish to express my sincere appreciation to all who attended the conference and to those who extended their kind support to organize the event successfully. I would like to express my deepest appreciation to the Panel of Judges, Professor Rohan Lucas, Mr D R Pullaperuma and Dr. Chulantha Kulasekara, the paper reviewers Prof Rohan Munasinghe, Dr Indika Perera and Dr. D. A. Meedeniya from University of Moratuwa, Professor Nihal Kodikara from University of Colombo and Dr. Pradeep Abeygunawardhana from Sri Lanka Institute of Information Technology, and the sessions Chairmen, Eng. Tharaka Mohotti and Eng. Rukmal Jayasinghe for their fullest support. The success of the Annual conference would not have been possible without the support of the organizing committee, i.e. Eng. Anuradha Jayakody, Eng. Dayan Gunawardena, Eng. Dasitha Liyanage and Eng. Naleen Jayasooriya. Last but not Least, special thanks go to authors who have done extensive research, written papers and made presentations.

I wish you all a very successful conference and hope to see you all at the next year's conference.

Eng. Dr. Lalith Liyanage
PhD (Northumbria/UK), MBA (eGov, Moratuwa/SL), MSc (Comp St, Essex/UK),
MSc (ElecEng, USSR), BSc (ElecEng, USSR), PG Dip (IIPM/India)
FIET(UK), MBCS(UK), MCS(SL), AMIE(SL)

Chairman – Conference Committee
23rd Annual Conference
IET – Sri Lanka Network

Message from the Editor

I would like to extend a warm welcome to the Engineering and Technology professionals and delegates on behalf of the publication committee, to the 23rd Annual Technical Conference of the Institution of Engineering and Technology (IET), Sri Lanka Network.

This year, the papers were mainly received in the fields of Electrical, Electronic, Telecommunication, Civil, Mechanical, Automobile, Marine, Aeronautical, Manufacturing and Information Engineering. The papers that were received are based on the individual, or collaborative research carried out in the industry and academia. After a careful review, Fourteen (14) high standard technical papers on various engineering aspects have been selected for presentations and publication out of 34 overall submissions. The review panel for the 23rd Annual Technical Conference of IET Sri Lanka Network 2016 proceedings considered not only the theoretical aspects of the papers but also the applicability of research outcomes of the papers for the industry. I would like to express my sincere gratitude to all the revivers, Prof Rohan Munasinghe, Dr. Indika Perera and Dr. D. A. Meedeniya from the University of Moratuwa, Professor Nihal Kodikara from University of Colombo and Dr. Pradeep Abeygunawardhana from Sri Lanka Institute of Information Technology.

I would like to express my sincere appreciation to all authors of the papers in this proceedings. It is the result of their generous contribution of time and effort on engineering and technology-related research. The willingness to make an effort to share knowledge and thoughtful insights with the engineering and technology community is greatly appreciated which has made this conference proceedings possible.

Finally, the organizers would like to thank all participants for their contribution, hoping that future IET annual Technical conference meetings will further bond synergies between the different research activities.

Eng. Anuradha Jayakody
B.Sc (Hons.)(Greenwich),(MSc (IS),PGD(IT),CITP ,MIEEE,MIET,MBCS,CCNA,CCAI,MCP)

Chairman – Editorial Committee
23rd Annual Conference
IET – Sri Lanka Network

Table of Contents

Androsafe: Online malware analysis with static and dynamic methods

Krishnadeva Kesavan, Chethana Liyanapathirana , S.A.W.S Sampath , Y.M. Sureni Koshila,Chamod Premarathne ,Sahan Wanigarathna, Chamira Priyamanthi Nawarathna, Prabhath Lakmal Rupasinghe

ABSTRACT

With an estimated market share of 70% to 80%, Android as becoming the most popular operating system for smartphone and tablet. Cyber criminals naturally expanded their various activities towards Google's mobile platform.An additional incentive for mobile malware authors to target Android instead of another mobile platform is Android open design that allows users to install the application from a variety of sources. "Androsafe" is an online malware analysis tool which can analyze malware in an isolated environment without any damaging to the mobile device by using both existing and new anomaly based and behavioral analysis. Through this combination, we can analyze a large number of malware families because some malware families may only perform signature base or behavioral. Then the sandboxes based on signature will not have analysis malware families that only perform a behavior and the sandboxes based on behavior will not analysis signature-based malware families."Androsafe" sandbox will be hosted in the Honeynet Project's cloud. Dynamic Analysis will be queued and run in the background, and an email which contains malware analyzing report will be sent to the user when the analysis is over. This method is very efficient more than offline kernel and app base sandbox.

1. INTRODUCTION

In the past couple of years, mobile devices have become sophisticated computing environments with increased computing power and network connectivity.

Android is a modern mobile platform that was designed to be open. Android applications make use of advanced hardware and software, as well as local and served data, exposed through the platform to bring innovation and value to consumers. To protect that value, the platform must offer an application environment that guarantees the security of users, data, applications, the device, and the network.

This open design easily allows attackers to perform attacks, such as social engineering attacks to mislead device users to install malware and attack third-party applications on Android. So we need to secure Android mobile device from this kind of common attacks. Securing an open platform requires strong security architecture.

When Android users download android applications from App stores to the mobile device, the user cannot test whether the application is secure or not. There can be hidden malicious codes inside the applications. To avoid such situations, we can test downloaded applications using sandboxing technology. Those sandboxes check applications for malware in offline mode. There's a risk behind that because offline checking can be harm to the mobile device. Attached malicious codes can breach the boundaries of the sandbox while analyzing the downloaded applications for malware.

There are two kinds of malware analyzing methods. One is a static method, and the other one is a dynamic method. Sandboxes currently available are based on either static or dynamic method. Then the analyzing scope will be less or limited. To overcome this limitation, we use both static and dynamic techniques to analyze malware.

We allow users to analyze downloaded applications by using the online tool "Androsafe." This reduces large impact over mobile devices by redirecting to an isolated environment apart from the mobile device to test downloaded applications, and there won't be any harm to the mobile device.

2. RESEARCH OBJECTIVE

The **outcome** of the research is an online Android sandbox "Androsafe" which is used to test suspicious programs that may contain a virus or other malicious code, without allowing the application to harm the mobile device.

We develop "Androsafe" to detect malware inside the downloaded applications. This online android

application sand box is capable of analyzing malware based on both static and dynamic methods.

Among static and dynamic malware analyzing tools, we decide to select Androguard as a static analyzing tool and Droidbox as a dynamic analyzing tool. Then we develop an Androguard algorithm inside the Droidbox algorithm in order to produce a combined module.

We first analyze the Androguard algorithm and find the malware analyzing techniques. The existing Androguard is able to detect only less number of malware families, and most of the time Androguard cannot detect the entire malware family. To address that problem we will develop the existing algorithm to detect those unknown malware families. Then the "Androsafe" algorithm is developed by inheriting the features from the "Droidbox" to the above-mentioned modified "Androguard" algorithm.

Droidbox will do dynamic analysis, and this will check for the behavior of malware. Droidbox is a tool to analyze Android apps. However, it lacks support to track native API calls. In fact, the current dynamic analysis methods running out of a method to track the native API calls during dynamic analysis. So we introduce the native API calls tracking mechanism to the current Droidbox algorithm.

We provide an isolated workspace in "Androsafe" to test any APK as an example chatting applications, gaming applications, etc. based on Android platform.

3. METHODOLOGY

Androguard is a python tool which is using static analysis method to detect malware in Android devices. Currently, Androguard detects several malware families such as Obad, Geinimi, DroidKungFu, Bracebridge. However, Androguard is lagging behind in detecting all the malware families such as FakeInstaller, etc. Our target is to enhance the current androguard algorithm to identify all malware families.

Androguard works with:

- Dex/Odex Dalvik virtual machine, .dex disassemble DE compilation.
- APK Android application.
- Android is binary XML.
- Android Resources.

Androguard has the following features:

It Maps and manipulates DEX/APK format into full Python objects.

- Disassemble/Decompilation/Modification of DEX/APK format.
- Decompilation with the first native directly from Dalvik byte codes to Java source codes Dalvik decompiler.
- Access to the static analysis of the code basic blocks, instructions, permissions and create analysis tool.
- Analyze a bunch of Android apps.
- Diffing of Android applications.
- Check if an Android application is there in a database.
- Open source database of Android Malware
- Reverse engineering of applications
- Transform Android's binary XML like AndroidManifest.xml into classic XML.

Basically, in Androguard, reverse engineering process is following the below steps.

1. Analyzing the Android-Manifest for permissions and activities

2. Unpacking of the Android application (apk file) to get all files and especially the classes.dex

3. Translating the Dalvik-Bytecode to Java-Bytecode

4. Analyzing the generated code

However, the problem is when Androguard gets the APK and try to unzip the APK, some malware generates anti-debugging tricks to crash Androguard. To overcome this problem we try to develop an algorithm to detect this kind of malware with anti-debugging tricks.

DroidBox consists of two parts, which can be referred to as the Host and the Target. The Target part launched on the emulator that is based on Android to monitor the data in low level.

The Host part is a set of Python scripts that connect with the emulator and receives all possible information from the Target regarding the application being analyzed and displays it in text or graphic format. DroidBox will output its results as a

JSON file.

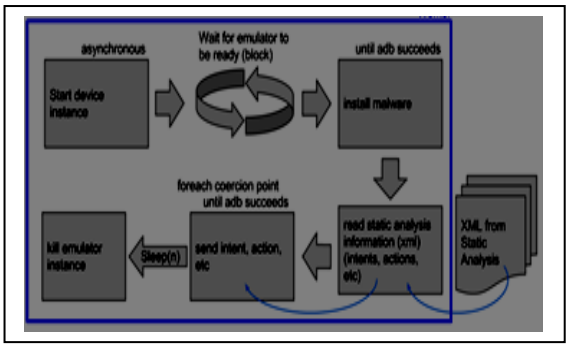


Figure 1: DroidBox Architecture

Dynamic analysis can monitor an APK's behavior utilizing following techniques in Droidbox.

- Taint tracking: Taint tracking tools are used in the dynamic analysis to detect potential misuse of users' private information.

- Virtual machine introspection (VMI): This intercepts events that occur within the emulated environment. Dalvik VMI-based systems monitor the execution of Android APIs through modifications in the Dalvik VM. Qemu VMI-based systems are implemented on the emulator level to enable the analysis of native code.

- System calls monitoring: Frameworks can collect an overview of executed system calls, by using, for instance, VMI, "strace" or a kernel module. This enables the tracing of native code.

- Method tracing: Frameworks can trace Java method invocations of an Apk in the Dalvik VM. It looks for malware family samples which are not detected by existing Droidbox.

According to the research done based on detecting Android Malware on Network Level in 2011, they found out several malware that are detected and not detected.

Not Detected Malware:

- Flexispy – Trojan horse family
- lovetrap - Trojan horse family
- kungfu – DroidKungFu family
- droiddeluxe – Droiddeluxe family
- basebridge – Basebridge malware family
- ggtracker - ggtracker family
- netisend and droiddream.
- Spygold and zone executed,

Table I – Detected Malware

Name	Identifying information
Crusewind	crusewind.net, HTTP-text, IMEI
walka	incorporateapps.com,
Tonclank	searchwebmobile.com, http-json IMEI
Bgserv	www.youlubg.com, IMEI, phone number
Smspacem	biofaction.no-ip.biz, HTTP-soap phone number
Lovetrap	cooshare.com, http-text positionrecorder.asmx, IMSI
(DL/installer)	api.go108.cn, http-xml no-store no-cache, love more and more

4. RESULT AND DISCUSSION

We are planned to get two sets of malware samples from different malware families for dynamic and static analysis separately. Then we analyze one sample using Androguard, and another sample will analyze using Droidbox. So far we have identified several malware samples from different malware families were not detected by Droidbox. We have considered malware samples from Trojan horse family, DroidKungFu family, Droiddeluxe family, Basebridge malware family, ggtracker family, netisend, and droiddream. Now we are planning to move this samples to check whether Androguard analyze them or not.

We use malware free Android applications from the third party app store called "China App store" to analyzing purpose. As well as we will maintain a malware signature database for static analysis. New Signature findings will send and store in the signature database. We use "Santoku 0.5" virtual machine in Linux Operating system.

We develop "Androsafe" over Linux platform and nowadays we are developing "Androsafe" web application which is used to submit the unverified APK file by users. At the same time, we are identifying useful and essential features that we can add from Androguard to Droidbox and redirect them to advance online malware analysis tool called "Androsafe."

"Androsafe" generates a report including personal or sensitive device information, Security appropriate actions, and logs regarding the access to hardware modules or sensors and communication information. We analyze, does the app try to access to the local address book, local calendar, stored pictures, configured accounts, local SMS or MMS messages, device identifiers and SIM card identifiers as personal or sensitive information.

As well as check security relevant actions like does the app use Crypto, load external libraries, try to modify device settings or try to install additional apps. We evaluate hardware modules or sensors deeply whether the app tries to use the camera, microphone, try to locate the device using the GPS sensor or network triangulation and does the app communicate with the Internet or the cloud services, Does the app try to send SMS messages, try to start a phone call, or try to open local ports.

5. CONCLUSION

Android is a free and open platform. This is also an advantage for developers as well as users. However, this will redirect many more security causes. This openness is a blessing as well as a curse for Android users. Anyone can put their Android applications in Android Marketplace. We cannot trust these applications are actually from developers or is there any hidden malicious code behind these third-party applications before install to our mobile devices. Most existing malware analyzing mechanisms are Kernel-based. In these mechanisms, the downloaded applications are checked for malware inside the mobile device, and it is very dangerous because sometimes infected malware can be activated inside the mobile device while analyzing and it can corrupt the entire system.

In our research, we develop an online Android application sandbox "Androsafe" by combining enhanced Androguard and Droidbox inside Droidbox to detect a significant amount of malware within an isolated environment hosted on a server. It is used to test unverified programs (APK) that may contain a virus or other malicious code, without allowing the application to damage the mobile device. Finally "Androsafe" generates an analyzing report including a summary of finding malware inside the downloaded APK. "Androsafe" is user-friendly, convenient, freely available online malware analyzing tool that anyone can access without any fear instead install downloaded Android applications blindly to your mobile device.

6. FUTURE WORK

This research is mainly focused on Android platform, but in the near future, it can be deployed for other mobile platforms. Currently, we are developing "Androsafe" web application, after that in the second step; we are supposed to develop an Android application. Our web application is developed in a way that any audience can use it with minimal effort. Also, we hope to develop this kind of features inside the "Androsafe" mobile application as well. We are aware of Malware and will develop "Androsafe" to identify the malware within the application. We are providing a user-friendly environment to the users in order to report any issues regarding the application like problems regarding malware. We are going to maintain a separate blacklist and whitelist. After we test the reported applications of users, we are going to group the applications separately. Applications with malware and bugs are labelled as blacklist applications, and the applications without any unauthorized modifications are labelled as whitelist applications. Users can download the application from our whitelist in the future. This will reduce the risk of downloading infected malware applications.

7. REFERENCES

[1] J. Bergeron, M. Debbabi, J. Desharnais, M. M. Erhioui, Y. Lavoie, and N. Tawbi. Static detection of malicious code in executable programs. In Proceedings of the Symposium on Requirements Engineering for Information Security (SREIS'01), 2001.

[2] A. Moser, C. Kruegel, and E. Kirda. Limits of static analysis for malware detection. In Proceedings of the 23rd Annual Computer Security Application Conference (ACSAC), pages 421–430, 2007.

[9] C. Willems, T. Holz, and F. Freiling. Toward automated dynamic malware analysis using cwsandbox. IEEE Security and Privacy, 5(2[3] P. Szor. Virus Research and Defense. Addison-Wesley, 2005.

[4] A.-D. Schmidt, J. H. Clausen, S. A. Camtepe, and S. Albayrak. Detecting Symbian os malware through a static function call analysis. In Proceedings of the 4th IEEE International Conference on Malicious and Unwanted Software (Malware 2009), pages 15–22. IEEE, 2009.

[5] Thomas Bl¨asing, Leonid Batyuk, Aubrey-Derrick Schmidt, Seyit Ahmet Camtepe, and Sahin Albayrak “An Android Application Sandbox System for Suspicious Software Detection”, in Technische Universit¨at Berlin 2010

[6] A. Desnos and G. Gueguen, “Android: From Reversing to Decompilation,” in Black Hat Abu Dhabi, Dec. 2011.

[7] S. Forrest, S. Hofmeyr, and A. Somayaji. The evolution of system-call monitoring. In ACSAC ’08: Proceedings of the 2008 Annual Computer Security Applications Conference, pages 418–430. IEEE Computer Society, 2008.

[8] M. A. Bishop. The Art and Science of Computer Security. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2002.

[10] A. Dewald, T. Holz, and F. Freiling. Moreover, sandbox: Sandboxing javascript to fight malicious websites. In Symposium on Applied Computing (SAC) 2010, Sierre, Switzerland, March 2010.

[11] T. Raffetseder, C. Kruegel, and E. Kirda. Detecting system emulators.

[12] M. Becher, F. Freiling, and B. Leider. On the effort to create smartphone worms in windows mobile. In Information Assurance and Security Workshop, 2007. IAW ’07. IEEE SMC, pages 199–206, 20-22 June 2007.

[13] Bundesamtf¨urSicherheitin derInformationstechnik. Mobile endger¨ateund mobile application: Sicherheitsgef¨ahrdungenund schutzmassnahmen, 2006.

[14] W. Enck, M. Ongtang, and P. McDaniel. Understanding Android Security. IEEE Security and Privacy, 7(1):50–57, 2009.

[15] S. Forrest, S. Hofmeyr, and A. Somayaji. The evolution of system-call monitoring. In ACSAC ’08: Proceedings of the 2008 Annual Computer Security Applications Conference, pages418–430. IEEE Computer Society, 2008.

[16] A. Rubini. Kernel system calls. http://www.ar.linux.it/docs/ksys/ksys.html. [Online; accessed 01-March-2010].

[17] GSM Association. IMEI allocation and approval guidelines, 2010. Available online at HTTP:// www.gsmworld.com/documents/DG06_v5.pdf; visited on December 4th, 2011.

[18] Eric Chien. Motivations of recent Android malware. Available online at http://www. symantec.com/content/en/us/enterprise/ media/security_response/whitepapers/ motivations_of_recent_android_malware. pdf; visited on December 4th, 2011.

[19] Gerry Eisenhaur, Michael N. Gagnon, Tufan Demir, and Neil Daswani. Mobile malware madness and how to cap the mad hatters, 2011. Available online at https://media. blackhat.com/bh-us-11/Daswani/BH_US_ 11_Daswani_Mobile_Malware_Slides.pdf; visited on December 4th, 2011

[20] Google management discusses q3 2011 results, 2011. Available online at http://seekingalpha.com/article/ 299518-google-management-discusses q3-2011-results-earnings-call-transcript; visited on December 4th, 2011

[21] B. Stone-Gross, M. Cova, L. Cavallaro, B. Gilbert, M. Szydlowski, R. Kemmerer, C. Kruegel, and G. Vigna. IMEI allocation and approval guidelines, 2009. Available online at http://www.cs.ucsb.edu/~seclab/ projects/torpig/torpig.pdf; visited on December 4th, 2011

Database Optimization Model to Reduce Application Delay Relates to Mobile Devices

J.A.D.C. Anuradha Jayakody[1], D.S.A.Kandawala[2], K.D.D Chathurangi[3], N.T Mapa[4], T.S Dissanayake[5]

Department of Information Systems Engineering, Sri Lanka Institute of Information Technology (SLIIT), Malabe, Sri Lanka.

[1] j.c.jayakody@postgrad.curtin.edu.au, [2] sachini.k@sliit.lk, [3] dinusha.c@sliit.lk, [4] navoda.m@sliit.lk, [5] thushari.d@sliit.lk

ABSTRACT

In the era of information technology, the smartphone has become an integral part of human lives. A variety of smartphone applications is available to provide different types of services required by the users. In order to execute installed applications, the mobile devices consume more memory and processing power. Hence to collect, store, transfer and combine massive amounts of resources. As lifestyles are becoming fast-paced, a slow and inefficient database applications are be deemed unacceptable. However, exploiting the information contained in these archives intelligently turns out to be relatively complicated, and it becomes harder to work with these data when it starts to evolve. It will be almost impossible to handle or access a massive amount of data, so it requires optimization method to reduce access time. This paper will discuss vital matrices such as access time, insert time, and delete time and execution time while increasing the performance of the database. Authors will focus on utilizing an Optimize indexes using Profiling to capture cost required to manipulate data. The outcome of this research work will improve the efficiency of the applications by minimizing the response time. The test results emphasize a comprehensive evaluation based analysis related to the proposed model.

Keywords **- database optimization, query optimization, indexing, profiling.**

1. INTRODUCTION

As lifestyles are becoming fast-paced, a slow and inefficient database applications are be deemed unacceptable. It will be almost impossible to handle or access a large amount of data. During this paper, vital importance is rendered in parliamentary procedure to address such problems that contribute to customer dissatisfaction while increasing the performance of data retrieval from the database [1].

Mobile applications have empowered and extended the usability of mobile devices far beyond merely supporting voice communication. The development of mobile applications, however, must deal with a variety of unique problems: limited working memory, limited storage, and limited processing power. Data storage is indispensable with the database. Hence, optimization on the database is the basis for application system to operate stably with high efficiency. Mobile applications which rely on remote data sources and databases are particularly challenging given the need to transmit data through wireless media and often involve complex business logic. Our main goal is to improve the performance of mobile applications which rely on remote data sources and databases. In this research work, we compare different data transmission optimization techniques, various middleware approaches and identify combinations of approaches for improving the performance of data transmission over the wireless network. The results of this research provide useful guidelines for the development of mobile applications needing to connect to remote databases or data sources.

This paper is organized as follows: Section I discuss the introduction to the area of research. Section II illustrates work related to this research area. System overview and the communication channel are comprehensively elaborate in chapter iii and IV

orderly. The methodology of the conducted research is explained in Section V with research findings. Finally, the conclusion and the future works are stated in the final part stated as Section VI.

2. LITERATURE SURVEY

This research work is a part of a larger project which is an implementation of an indoor navigation solution called "Third Vision," an android application integrated with beacon network and localization technique. The following sections describe the literature review that authors considered to increase the performance of the data retrieval and database optimization. Techniques used to get the data between Android application, and database system is REST (Representational State Transfer), client implementation and development inside the Android application, using HTTP methods to set and get information from the application and database, and JSON format to read and transfer information in an easy and understandable way, both from the Android application and the database. REST allows better support for browser clients due to its support for JSON. REST has better performance, scalability; reads can be cached, and JSON feeds can be loaded asynchronously much more easily than XML/RSS [2].

According to the authors, considerations following drawbacks are identified as a responsible for the low performance of MySQL databases. After optimizing the hardware, the operating system and then the SQL server settings, the main factors which affect the speed of execution are missing indexes, Inexact statistics, badly written queries, excessive fragmentation of indexes, frequent recompilation of queries. By understanding how the SQL Server optimizer works is the first step toward the establishment of a truly optimized database environment. These are only a few of the factors which can negatively influence the performance of a database. To investigate solutions for the problems mentioned above, need to research this area.

2.1 Indexing

This technique is one of the most beneficial methods for increasing the performance of queries is the creation of efficient indexes. A well-constructed index will allow a query to avoid the necessity of scanning the entire table for results. While the B-tree (or the B+ tree) is the most popular index structure in disk-based relational database systems, the T-tree has been widely accepted as a promising index structure for main memory databases where the entire database (or most of them) resides in the main memory. However, most works on the T-tree reported in the literature did not consider concurrency control. As a solution, the primary key should be created for each table. Once a primary key set up in a table, a clustered index dictates the storage order of the data in a table. Clustered indexes are more efficient on columns of data that are most often searched for ranges of values. All data pages containing the table rows are physically sorted in the file system according to their primary key values. A non-clustered index is similar to an index in a textbook where the data is stored in one place and the data value in another [3]. The evaluation for any of the index must consider the following factors.

- ***Access Type:*** It can effectively support the access type, according to the specified attribute value to find the appropriate record and according to the range of the attribute values to find all records in this range.

- ***Access Time:*** The time that is spent to access one or more data.

- ***Insert Time:*** The time to insert a new data in the index including the time of finding the correct location and the time of correspondingly modifying the index structure.

- ***Delete Time:*** The time to delete a data in the index including the time of finding the location of the data that will be deleted and the time of modifying the index structure.

Each data page contains rows which are also sorted within the data page according to their primary key values. So, each time we ask any row from the table, the database server finds the corresponding data page first using the clustered index tree and then finds the desired row within the data page that contains the primary key value. Following is how an index tree looks like. At present, there is T tree and B-tree in the index of the memory database system design and implementation [14], [15], [16].

2.2 Statistics

As a general rule, indexes should be kept as narrow as possible, most likely following a business use requirement, to reduce the amount of processing

overhead associated with each query. Analyzing and optimizing query performance requires the measurement and collection of statistical data. SQL Server maintains statistical performance data about index keys. If configured to do so, non-indexed keys can also be statistically measured. There are several basic statistical concepts to understand regarding performance optimization.

- ***Cardinality:*** Measures number of unique values exists in the data set.

- ***Density:*** Measures the uniqueness of values within a data set. Density is determined by dividing the number of rows that correspond to a given key by the number of rows in the table. The optimizer will likely ignore indexes with high densities.

- ***Selectivity:*** Measures the number of rows that will be returned by a particular query. Selectivity is determined by dividing the number of keys requested by the number of rows they access. To select the valid measure of the cost, the optimizer should calculate a necessary query plan [4].

2.3 Appropriate covering indexes

If we know that our application will be performing the same query over and over on the same table, we should consider creating a covering index on the table. A covering index, which is a form of a composite index, includes all the columns referenced in SELECT, JOIN, and WHERE clauses of a query.

2.4 Data Profiling

Data profiling comprises a broad range of methods to analyze a given data set efficiently. "Data profiling" is the process of examining the data available in an existing data source and collecting statistics and information about that data [17], [18].

- ***Incremental profiling***: An obvious, but yet under-examined extension to data profiling is to reuse earlier profiling results to speed up computation on changed data.

- ***Continuous profiling:*** If the profiling results can be expressed as a query, and if they shall be performed only on a temporal window of the data.

- ***Multi-measure profiling:*** Each profiling algorithm has its scheme of running through the data and collecting or aggregating whatever information is needed. Realizing that multiple types of profiling metadata shall be collected, it is likely that many of these runs can be combined [19], [20].

2.5 Query Optimization and Indexing

According to the Fan Yuanyuan, MiXifeng on "Distributed Database System Query Optimization Algorithm Research" introduces a new algorithm. It is designed through the research on query optimization technology, based on a number of optimization algorithms commonly used in the distributed query, a new algorithm is designed, which can significantly reduce the amount of intermediate result data, effectively reduce the communication network cost, to improve the optimal efficiency [5]. This is research on producing an algorithm to improve the efficiency of the database by using query optimization, so the research is not focusing on the use of indexing.

In the paper of Lin Hong, Zhuhai, Mingda Lu and Weiting Hong on "A Business Computing _System Optimization Research on the Efficiency of Database Queries", they discussed the performance evaluation, measurement, and business computing system optimization based on some experiment researches on the efficiency of database queries [6]. And not considering the use of indexing.

In the paper of Ivo Jimenez, Jeff LeFevre et-al. On "Benchmarking Online Index-Tuning Algorithms", they outline the development of a performance benchmark for the specific problem of online index tuning [7] and not considering about the query optimization techniques.

In the paper of Surajit Chaudhuri, Vivek Narasayya on "SelfTuning Database Systems: A Decade of Progress", they discuss advances in self-tuning database systems over the past decade, and it primarily focuses on the problem of automated physical database design [8].

From the paper of P. Karthik, Prof. G. Thippa Reddy and E. Kaari Vanan discussed on "Tuning the SQL Query in order to Reduce Time Consumption". They have presented a new query optimization method which gives the High performance of the system and less stress on the database when data transmission occurs and the efficient usage of a database engine with reduced memory requirements [9].

According to the literature reviews carried out, which is encouraging for tuning the databases to

increase the performance and reduces the latency of mobile application. Still, there is a necessity for a standard approach to increase the performance of indexing and query optimization techniques. This paper focuses to design a model and evaluate it in a real-time environment.

3. SYSTEM OVERVIEW

The application can be accessed by anyone, the interaction between the application and the user is based on voice and the gestures. After user gives input to the application, it responses to the user with the help of an Optimized database. (Figure. 1) Depicts the system overview and research paper focuses area.

Figure.1: High-Level Architecture of the System

4. COMMUNICATION CHANNEL

As depicted in (Figure. 2) REST is used to increase the performance of the data retrieval.

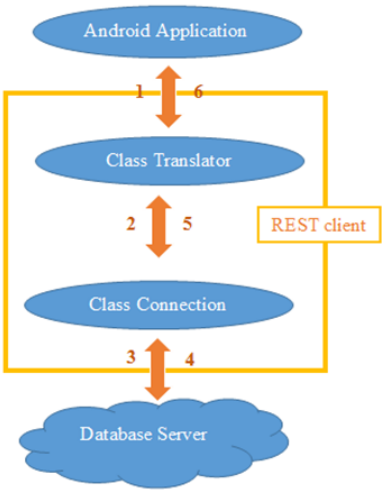


Figure .3: REST Client-Server Architecture

Steps of the above REST operation can be illustrated as follows.

1. The information is sent from the Android application to the translator class.
2. The translator class transforms the information to JSON-format and transfers it to connect class.
3. The connect class makes the connection to the database server and sends the information.
4. If the connect class requests an answer from the database server, it is sent back in JSON-format.
5. The information is sent back to the translator class and extracted. The translator class checks where to send the information inside the application before it is sent.
6. The information is used inside the application.

5. RESEARCH METHODOLOGY

Mobile Application delay directly concerns with the data access through databases. Hence to control and perform the actions related to database access efficiently, the database must be well optimized with corresponding rules and principles. Even though there are verities of optimization methods available to overcome the problem, there is a necessity to provide a convenient solution when it comes to mobile applications. Database performance tuning core areas are proposed to apply to this mobile application delay issue, with the help of Indexing and Query optimization methods.

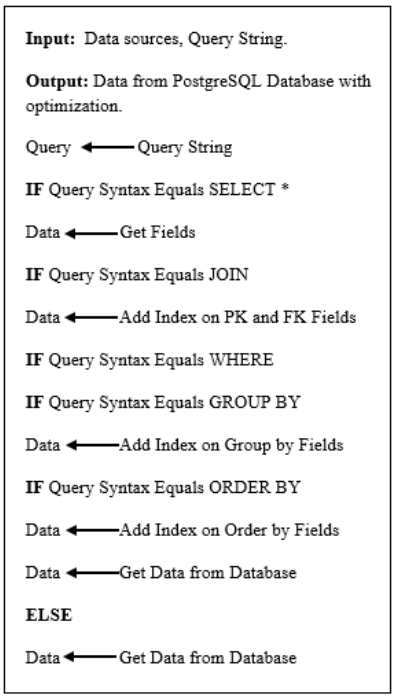


Figure .4: Pseudo code of proposed database optimization approach

As a solution for this research problem, a standard Database Optimization Approach was introduced; this will provide a specific standard to follow up when applying optimization techniques mostly related to query optimization and indexing. The Proposed database optimization approach has illustrated using indexing and query optimization methods in (Figure.3). On the model, two queries have produced to test two main cases in order to evaluate the proposed database optimization model which is shown in (Figure.4). Cases can be sequentially discussed as follows.

- ***Case 1:*** Tested results for the Proposed Database Optimization Model

- ***Case 2:*** Tested results with a Non-Optimized query.

With respect to above-mentioned cases, the query can be generated as follows.

```
1. SET profiling = 1;
2. SELECT device.location, doctor.name, destination.path
        FROM device, doctor, destination
3. INNER JOIN    destination ON destination.id
                 =doctor.destination_id
4. INNER JOIN  doctor ON doctor.id = doctor.device_id
      WHERE device.mac_address =" 01:23:45:67:89: ab";
5. SHOW profiles;
```

Figure .5 The query used to evaluate the proposed two test cases

According to the query specified in above figure (Fig.4), every query line can be described in order to satisfy the standard database optimization approach (Figure.3).

1. Enable profiling by setting it to 1.
2. Select the device location, doctor name, and destination from the device, doctor, and destination.
3. Join both the doctors and destination table using the corresponding ids.
4. Join both the doctor and device table using the corresponding ids, where mac addresses equal to = “01:23:45:67:89: ab”.
5. Display a list of the most recent statements sent to the server.

To prove that it is efficient and practical in real-time environments. These queries have been tested and simulated using the PostgreSQL, and it executed on Intel® Core™ i7, 8GB RAM. In order to test above two cases, 2000 records have been used from three main tables, namely, Device table, Doctor Table and Destination table.” Third Vision”, a Smartphone based indoor navigation application has been used to evaluate the proposed approach in a real-time environment.

The test results were realized concerning the number of queries that were executed together by the database. Hence in each case, the test results were taken while changing this factor (number of queries executed at the same time) from 1 to 20. Profiling has been used to measure the execution time of the queries and database selection time that was executed concurrently. Mostly the query execution time is a very small value which comes in milliseconds. In order to attain a considerable difference between these two tested cases of each situation, the frequency has been calculated by taking one over execution time.

Frequency = 1/Execution Time $\rightarrow$ (1)

Table I – Test Results for the comparison between two cases

No.of occurrences (x)	Test result with query proposed optimization procedure Frequency (y1)	Non-optimized query Frequency (y2)
1	3024.25	2289.42
2	2566.4	1951.81
3	2075.7	1035.22
4	1220.51	876.76
5	768.28	590.2
6	637.12	483.25
7	560.66	445.66
8	510.13	316.28
9	425.48	244.96
10	390.26	198.01
11	345.64	165.73
12	327.46	140.61
13	297.2	123.37
14	245.39	113.42
15	199.2	103
16	156.1	96.47
17	122.9	64.2
18	99.4	41.09
19	43.25	23.45
20	10.6	8.11

Based on the test results obtained from the database, a chart can be plot as follows to view the significance of the proposed approach. According to the chart shown in (Figure.5), the x-axis of the above chart displays the number of queries that were executed at the same time by the database (no. of occurrences). The y-axis displays the frequency of the relevant query set. This graph interprets that the proposed optimization approach has a high slope than the non-optimized query, and also it takes low execution time to execute the query.

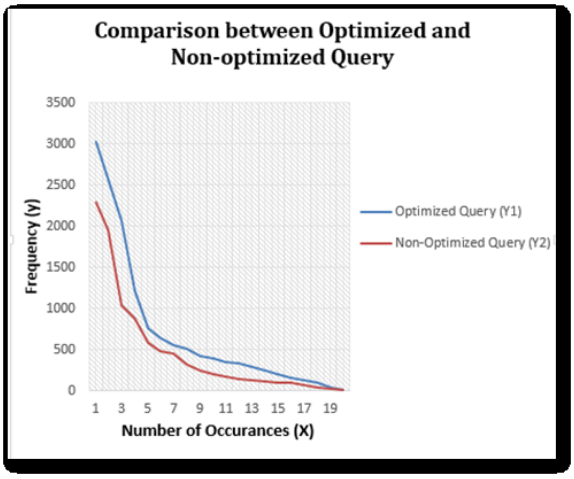


Figure .6: Comparison between two queries

Based on these evaluations it is confirmed that the proposed Database Optimization Approach provides an accurate, efficient and effective solution for the identified research problem.

6. CONCLUSION & FUTURE WORK

This research work demonstrated that the proposed model gives a unique opportunity to apply to the real scenarios in the case of data retrieved from the database using the SELECT command. Experimental results show that the requirement of the index, query optimization for the database and use of REST protocol techniques are needed to speed data retrieval.Further discussed other optimization techniques like Statistics, Data profiling, and Query Optimizing. These will improve the efficiency of the application so that the response time is minimized. In the future work planned to implement the same procedure for INSERT, UPDATE and DELETE commands. This will assist a large number of data retrievals from database to application.

7. ACKNOWLEDGEMENTS

This work has been supported by the Sri Lanka Institute of the Information Technology, Malabe Sri Lanka.

8. REFERENCES

[1] Miran Kim, Myungsun Kim, “Optimized Search and Computer Circuits and their Application to Query Evaluation and Encrypted Data”, South Korea, in IEEE, 2016.

[2] REST vs SOAP, The Difference between Soap and Rest, online available, http://spf13.com/post/soap-vs-rest [accessed on 13.05.2015].

[3] W.Guo and Z.Hu, "Memory Database Index Optimization", Wuhan, China, IEEE, 2010.

[4] O.ChunLin, T. Kwee, and S. Tsai, “Database Optimization for Novelty Detection”, Macau, IEEE, 2009.

[5] Fan Yuanyuan and MiXifeng, “Distributed Database System Query Optimization Algorithm Research”, Department of Computer and Information Engineering, Jiaozuo Teachers College, 2010.

[6] Lin Hong et al., “A Business Computing System Optimization Research on the Efficiency of Database Queries”, in International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery, Int. Conf., 2013.

[7] Ivo Jimenez et al., “Benchmarking Online Index-Tuning Algorithms”, University of California Santa Cruz, 2011.

[8] Surajit Chaudhuri and Vivek Narasayya, “Self-Tuning Database Systems: A Decade of Progress”, Microsoft Research, 2007.

[9] P. Karthik et al., “Tuning the SQL Query in order to Reduce Time Consumption”, IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 4, No 3, School of Information Technology and Engineering, VIT University Vellore, Tamil Nadu, India, July 2012.

[10] D.Li, L.Han, and Y.Ding,"SQL Query Optimization Methods of Relational Database System”, Bali Island, IEEE, 2010.

[11] Ong Chun Lin et al., “Database Optimization for Novelty Detection”, School of Electrical & Electronic Engineering, Nanyang Technological University, Singapore, 2009.

[12] Maxim Martynov and Boris Novikov, “An Indexing Algorithm for Text Retrieval”, University of St. -Petersburg, Russia, 1996.

[13] “Data Profiling,” 2015. [Online]. Available: https://en.wikipedia.org/wiki/Data_profiling.

[14] LiWu Chang, Ira S. Moskowitz, James Tracy, "An Agent-based Approach to Inference Prevention in Distributed Database Systems".

[15] S.F. Rodd, Dr, U.P. Kulkarni, "Adaptive Tuning Algorithm for Performance Tuning of Database Management System,” 2010.

[16] Hitesh Kumar Sharma, Aditya Shastri, Ranjit Biswas, "Architecture of Automated Database Tuning Using SGA parameter", 2012.

[17] David J. Montana and Lawrence Davis, Training Feedforward," Neural Networks Using Genetic Algorithms".

[18] S. Agarwal and et al., “Automated selection of materialized views and indexes”, VLDB, 2007.

[19] K. Schnaitter, “On-line index selection for physical database tuning,” Ph.D. dissertation, University of California Santa Cruz, 2010.

[20] J.A.D.C.Anuradha Jayakody, Iain Murray, Johannes Herrmann "Database modelling for vision impaired indoor navigation system", ICT, 2015.

A Sentiment Analysis and Classification Algorithm Utilizing an Independent Term Matching Scheme Sensitive to Negations and Word Count Patterns

Dr. Asoka Korale, C.Eng. MIET, Chanuka Perera, Dip. ,ABE(UK) Eranda Adikari, BSc., C.Eng. MIESL, Nadeesha Ekanayake, B.Sc.

ABSTRACT

The "Sentiment" or "Emotion" contained in a short comment particularly when expressed by a customer in relation to a product or level of service provides valuable feedback to an organization on customer satisfaction enabling timely actions to resolve current and emerging issues of varying degrees of importance to the business. The ability to measure the "Sentiment" and particularly its "degree" allows the severity of the feedback to be quantified. Independent Term Matching Algorithms (ITMA) while allowing for the calculation of an overall sentiment score for a comment made up of several emotion bearing words or phrases typically do not take into account negations in the comment and therefore miss important meaning contained therein [3]. Typically these algorithms cannot also be trained to recognize patterns within the comments and so are unable to classify comments as to their inherent content.

In light of this, we propose a novel algorithm that allows rating multiple emotion is bearing words in a comment using the ITMA while also accounting for negations. This is effected by modifying the sentiment score of the emotion bearing word occurring subsequent to the negation so that its value is adjusted in a direction opposite in polarity to the affected word. The proposed model is also able to classify comments according to their content so they can be directed to the relevant authority for necessary action. This classification is vital as it allows an assessment of the severity of the issue via its sentiment score while also allowing steps for its resolution in a timely and effective manner by individuals with the relevant domain expertise.

We use a process of Association Rule Discovery to find groups of words that are likely to occur together in a comment to facilitate the classification. This allows the generation of word combinations or word sets that are used in training a Naive Bayes Classifier. The rules are selected based on support, confidence and coverage criteria so that only those that are meaningful to the classification are chosen. This process of keyword generation and association rule prioritization allows us to classify a diverse set of comments relating to a number of sources and a range of issues emanating from a varied group of customers concerning sentiment, its degree, and its content.

1. INTRODUCTION

Any Organization that has a client interface or touch point will invariably have to deal with customer requests and feedback on the offered products and services. These touch points may be in the form of Corporate Websites, FaceBook, and Social Media pages, Call Centers and regular Sales and Support outlets where customers will interface and interact with the organization either through its representatives or via electronic means. In every such instance, it is important for an organization interested in meeting its customer expectations to assess the degree of customer satisfaction by soliciting feedback. In fact, customer satisfaction surveys often form the cornerstone of organizational strategy providing indications as to where targets are being met and where improvements are needed in a wide variety of operational areas of the business usually focusing on its Product Offering and Service Quality.

This customer feedback is usually in the form of short comments and answers to specific questions put to the customer that is usually recorded in electronic form. The feedback that is received via

structured questionnaires having a choice of predetermined responses is rather straightforward to analyze and arrives at its conclusions using statistical methodology. The comments are however unstructured and may relate to any topic and any aspect of the organization the customer has interacted with and is thus more challenging to analyze and quantify.

The comments, however, are a richer source of information on what customers wish to convey to the organization. A means to successfully analyze this aspect of the feedback and distil the essence of what is conveyed provides an organization with valuable insights into its day to day operations and a way to map out its future strategies.

The sentiment or emotional content of the comments and the specific topic or areas to which it relates are the broad aspects that usually concern a business. Evaluating customer sentiment then allows an organization to assess its standing in the eyes of its customers and the public and when observed in the temporal dimension, enables valuable feedback as to how the company and its operations are viewed on a day to day basis. The time dimension is a key attribute as certain types of feedback and customer perceptions vary with time of day and day of the week as some issues are inherently time-bound and occur intermittently or with some regularity.

"Sentiment" can be defined as a view, an opinion held or expressed [4]. It is a representation of the attitudes, emotions, and opinions that is held by an individual and is concerned with impressions and not fact. Thus words bearing emotion convey the sentiment. The analysis of such words in isolation and groups enables us to quantify the emotional content inherent in a particular corpus. Detailed dictionaries have been compiled through large surveys conducted across different demographics, and the emotional content of words have been measured on numerical scales and tabulated. These dictionaries enable algorithms to analyze and rate the sentiment content and its degree of a sequence of words.

Sentiment Analysis is thus critical to organizational success in that it even influences how speculators trade in the stock market taking advantage of the emotional content of news items appearing in print and electronic media concerning different aspects of an organization and its products. It is not unusual for a company to monitor social media and other electronic channels to assess customer satisfaction levels on a new product or the public response to a particular course of action. Governments too are now in the habit of mining media channels for feedback on popular sentiment to its policies and leadership.

Mobile Telecommunications Service providers receive customer feedback from many of the physical and electronic touch points in the form of customer service centres, call centres, and social media pages. Service quality is particularly a key differentiator in this space making customer feedback a critical component of devising an organizational and operational strategy. In fact, Service Quality Issues raised from customer feedback is sometimes the only way in which Network performance issues are initially identified and remedial actions were taken. Feedback on service quality is also key business driver helping to determine where future investments in infrastructure are to be directed. The day to day performance of the network can also be monitored in this context. In light of this, the feedback comments received via the different channels and touch points must be directed to the appropriate authority for action soon as possible via some automated means.

These factors are then the principal drivers for an organization to employ Sentiment Analysis and Classification of customer feedback comments. In light of this, we propose a novel technique that analyzes short comments, rates their emotional content and classifies them into a set of predetermined categories based on a set of pre-classified comments used as a training sample.

2. GENERAL APPROACH AND FEATURES OF THE ALGORITHM

By this modelling, an organization is able to classify and direct comments to the relevant authority while assigning a sentiment score to each comment. The sentiment score of the comment is an indication of the severity of the customer feedback which also allows a certain priority to be assigned to each class of comment. The classifier is trained using word associations (rules) derived from Association Rule Mining of the comments prioritized by confidence level.

2.1 The proposed Sentiment Analysis Algorithm

Independent Term Matching [3] is a technique that is amenable to the analysis of short comments. The comments analyzed and modelled in this particular use case are typical of the length of one or two dozen words and so do not present a large corpus of text for analysis via methods that require training.

We enhance the Independent Term Matching algorithm by incorporating the facility to handle negations in the comment that change the emotional content of the word following the negation.

2.2 Approaches to the Classification Scheme

A classification scheme modeled on the Naïve Bayes technique is employed where a set of pre-classified comments are analyzed for word frequencies. In this approach, it is the occurrence of words and not the order of their occurrence in a particular comment that is important. In other words, the set of words in a comment are considered to be independent of each other given the particular class to which that comment belongs. The algorithm proceeds by determining the conditional probability of a word given a particular class and obtains the predicted class via Bayes theorem and the assumption of the conditional independence of each word given its class.

2.3 Association Rule Mining in the Classification

Association Rules provide an insight into the occurrence of words and word combinations in a comment. It acts as a guide in picking certain keywords that are useful in classifying the initial training comments.

In this modeling, we consider each comment as a transaction and the particular words in the comment as the items in that transaction. The Association Rules are then mined for the aggregation of all comments (transactions) which then is cast in the light of a traditional market basket analysis giving association rules (relationships) between words in a comment.

The rules are quantified via performance measures derived from Confidence and Lift to determine which words are most likely to occur together and provide a guide in determining the class of the comment based on the occurrence of certain keywords in the initial classification-training phase.

3. CURRENT STATE OF THE ART IN SENTIMENT ANALYSIS AND CLASSIFICATION SCHEMES

Existing Sentiment Analysis techniques usually classify the sentiment of a text into three subjective categories “positive”, “negative” and “neutral”. These techniques rely on a supervised learning algorithm that uses pre-classified texts to learn the patterns of words in comments that bear a particular category of sentiment. The classification result is only as good as the training samples offered to the learning algorithm to discover the emotional content of words and their grouping given a particular class of comment.

Given the wide variety of comments and words therein, the number of pre-classified samples would also have to be large in order to use this technique successfully in the context of our particular use case.

Independent Term Matching Algorithms like the one used in this paper has no recourse to a sequence of training comments but utilize ratings for emotion bearing words computed from surveys. In this regard, this technique is more amenable to handle short comments or when the amount of text is too small to train a model. A shortcoming of this technique is its inability to handle the effect of negations. This class of algorithm, however, can produce a continuum for the sentiment rating depending on the precision of which the words in the Dictionary are rated by combining the effect

While the Naïve Bayes “bag of words” technique is widespread, this classification typically does not use keywords generated via Association rule mining of the comments. In fact, the two techniques of Naïve Bayes and Association Rules can be considered complementary in that both look for patterns of words in a comment which are then aggregated across all comments to derive probabilities and rules respectively.

4. MAIN CONTRIBUTIONS OF THE PAPER

1. *A novel Association Rule Mining Algorithm*

A flexible technique designed to generate only the required set of rules at a particular level or depth with rules sorted and prioritized by confidence

level. The important rules provide insight into keywords that should be used in the classification of comments used in model training.

2. *Sentiment Scoring incorporating the effect of Negations*
The Independent Term Matching Algorithm is enhanced with the ability to rate the sentiment of words modified by preceding Negations giving a more complete, holistic and accurate rating of the emotional content of a comment. The sentiment of such affected words is adjusted in the direction opposite to its current polarity by an amount that reflects the uncertainty in its rated sentiment.

3. *Classification via Naïve Bayes technique utilizing Key Words derived from Association Rules*
Classification of comments via Naïve Bayes while an established method benefits from the added insights gained from the keywords used in the classification of the training data comment set. The use of word associations derived from the rules displaying high confidence allows a higher accuracy and greater consistency in the classification of the training data resulting in better overall performance in the prediction.

5. COMPONENTS OF THE PROPOSED ALGORITHM

Association Rule Mining, Comment Classification, and Sentiment Rating are the three principal aspects of the proposed model. Although Sentiment Analysis and Classification may be considered in isolation, our use case demands that we rate the sentiment of the comments in a particular predicted class for prioritization of handling.

5.1 Association Rule Mining

The Association Rules are mined from an Incidence Matrix (IM) by placing apriori limits on the support for each Item which is the column sum of IM. IM is formed as a comment vs. word (Item) matrix, by allocating unity for every word that is found in each comment. If a word appears in a comment more than once, the entry in the corresponding word column in the row allocated for that comment is still unity.

Once the column sum (IMsum_min) is formed the particular columns that do not meet the minimum support criteria (min_supp) are not considered and eliminated from the analysis as a rule that incorporates that item cannot, in any case, meet the minimum support requirement.

The rule discovery then proceeds to the first stage by determining all pairs of items that meet the minimum support criteria. The support for a pair of items is then simply the dot product of the corresponding columns of IM. In the next step, the pairs of items that meet the minimum support criteria are selected, and the support and confidence of the rule are also determined.

In the second stage, only the rules that met the minimum support criteria from the first stage are selected to form the rules. The antecedent in the second stage is then formed by merging antecedent and consequent terms from the rules (that met the minimum support criteria) of the first stage. The three-item rules (2 in items in the antecedent derived from stage 1, and one new item in the consequent) that can be formed that meet the minimum support criteria can only come from the set of terms selected from stage 1 and the set IMsum_min.

For example, let A, B, C, D be items that meet the minim support criteria in IM.
In the first stage all pairs {A,B}, {A,C}, {A,D}, {B,C}, {B,D}, {C,D} are tested for minimum support. (There will be 4C2 = 6 such terms to be tested.)

A->B is a valid rule, only if the support for AUB meets min_supp criteria

In the second stage A and B (and all other pairs that met min_supp) (are merged to form the antecedent of the rule at stage 2 if they meet the min_supp criteria. (let us assume that {A,B}, {A,C} and {C,D} meet the min_supp criteria)

Thus in stage 2
A&B->C is a valid rule only if the support for AUBUC meets min_supp criteria (and similarly we test support for rules A&B->D, A&C->B, A&C->D and C&D->A and C&D->B)

The rule discovery proceeds in this manner until all unique combinations of items have been created in the rules.

5.1.1 Definitions and Performance Measures

Support (A->B) = N(AUB)/N (1)

where *N*(AUB) represents the number of times items A and B are brought together in the whole transaction set and N is the number of rows in IM (corresponds to the number of unique subscribers or transactions)

Confidence(A->B)

$$= \text{Support}(AUB)/\text{Support}(A) \quad (2)$$

$$= P(E_A \& E_B)/P(E_A) \quad (3)$$

$$= P(E_B/E_A) \quad (4)$$

Where $P(E_X)$ is the probability of word (item) X being present in a particular comment.

The confidence of a rule (A->B) is a measure of the validity of the rule; it has a conditional probability interpretation in that, it is the probability that B occurs when A has already occurred. In market basket parlance it is the probability of word (item) B being in a comment given that A is in the same comment.

Loosely speaking, we use the confidence measure in our reasoning as an indication of the degree of correlation between words when we have rules of the form word1 -> word2. It is also used to rank the association rules particularly when one common antecedent gives rise to many consequent items.

The Lift is a popular measure used to measure how likely a rule is if the items are independent.

Lift(A->B)

$$= \text{Confidence}(A->B)/\text{Support}(B) \quad (5)$$

$$\text{Lift}(A->B) = P(E_A \& E_B)/P(E_A)P(E_B) \quad (6)$$

5.2 Sentiment Rating via Independent Term Matching while accounting for Negations

The overall sentiment of a comment is determined by aggregating the effect of the individual words that bear emotional content or sentiment. The words that bear sentiment are identified by the existence of a match in a dictionary of 14,000 words compiled by Warriner et al. This dictionary contains a mean and standard deviation value for each emotion bearing word in the dimensions of Valence (Happiness), Activation (Arousal) and Dominance. These values are the result of a survey of raters where it is assumed that the distribution of the ratings for each word follows a Normal Distribution. The means and standard deviations for each emotional dimension of each word provided in the Dictionary are then the means and standard deviations of the ratings given by the population of raters who rated each word.

Thus by aggregating the effect of the all such matched words in a comment by summing the corresponding Normal Probability Density functions corresponding to each matched word we are in effect treating the overall density of a particular sentiment dimension of a comment as a sum of Gaussians each with a particular mean and standard deviation. In effect, this is a Gaussian Mixture Model (GMM) where the overall sentiment rating of a particular dimension is then determined as the value that gives the highest probability of the cumulative sum of the Normal individual densities.

In this modeling it is assumed that each matched word is equally likely to occur in particular comment and that a particular emotional dimension of each word is independent of its other emotional dimensions. Thus we can consider the cumulative effect of all matched sentiment-bearing words via the sum of the individual probability densities.

$$f(x;\theta) = \sum_{k=1}^{N} p_k g(x; m_k \sigma_k) \ , \quad (7)$$

where x represents the sentiment score, N the number of matched words in a comment m_k, σ_k the mean and standard deviation of the Normal Distribution of the ratings of each matched word,

$$p_k = \frac{1}{N} \quad (8)$$

as the words are of equal probability of occurrence and are also independent of each other, and

$$g(x; m_k, \sigma_k) = \frac{1}{\sqrt{2\pi}\sigma_k} e^{-\frac{1}{2}\left(\frac{\|x-m_k\|}{\sigma_k}\right)^2} \quad (9)$$

The overall sentiment of a comment then is determined as

$$x_{comment} = \max_x f(x;\theta) \quad (10)$$

Which is the point at which the probability of the mixture of distribution is a maximum, and so is the most likely value for the overall sentiment of a comment composed of several words.

Thus a comment of the form "the service was good but was late" finds a match in the Dictionary for words "service", "good" and "late". The cumulative effect of the three words in the comment and therefore the overall sentiment of the comment in the Valence and Activation dimensions are displayed in Figure 1 and Figure 2. The Valence and Activation ratings for each word matched in the Dictionary and their cumulative effect via GMM is found in Table 1.

Table 1: Dictionary Sentiment Ratings for each matched word in the comment.

Comment Words	Valence Rating		Activation Rating	
Dictionary Value	Mean	Std Dev	Mean	Std Dev
'service'	6.83	1.54	2.95	2.09
'good'	7.89	1.24	3.66	2.72
'late'	3.32	1.17	5.57	2.56
Simple Average	**6.01**	**1.32**	**4.06**	**2.46**

Table 2: Cumulative Rating for entire comment via Gaussian Mixtures.

Word	Valence Rating	Activation Rating
max- GMM	**7.5**	**3.7**

Thus the overall sentiment ratings found via a maximum probability criteria is somewhat different to a simple average of the values. But in most instances the simple average and the max-GMM results are very close as short comments usually only contain words of a single polarity and do not in general feature words bearing emotions contrary to one another (like "late" vs. "good" and "service"). One can readily observe that the distributions for "service" and "good" are close to one another as they in general express similar sentiment. The presence of "late" in the comment alters the Gaussian mixture such that the Simple Average and the max-GMM value are at variance.

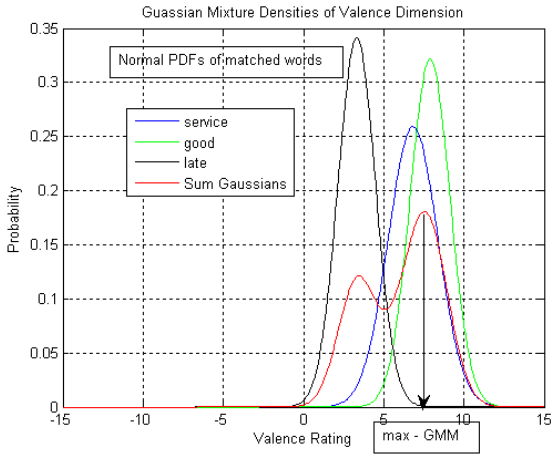


Figure 1: Gaussian Mixtures of matched words in the Valence Dimension

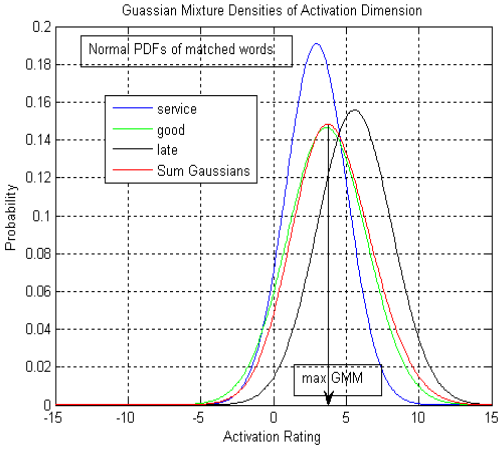


Figure 2: Gaussian Mixtures of matched words in the Activation Dimension

5.2.1 Accounting for Negations

In this paper, we also account for Negations that modify the meaning of word immediately following the negation by adjusting the sentiment score of the word immediately following the negation in a direction opposite in polarity to its matched directory sentiment value. The magnitude of the adjustment made corresponds to the standard deviation of the particular rating value being adjusted. Thus analysis of the comment "the service was not good and late" results in the following.

Table 3: Dictionary Sentiment Ratings for each matched word in the comment.

Comment Words	Valence Rating		Activation Rating	
Dictionary Value	Mean	Std Dev	Mean	Std Dev
'service'	6.83	1.54	2.95	2.09

not 'good'	6.65	1.24	6.38	2.72
'late'	3.32	1.17	5.57	2.56
Simple Average	**5.6**	**1.32**	**4.97**	**2.46**

Table 4: Cumulative Rating for entire comment via Gaussian Mixtures.

Word	Valence Rating	Activation Rating
max- GMM	**6.7**	**4.5**

Our reasoning here is that while "not good" is not the same as saying that it is "bad," it, however, has the effect of diminishing the positive sentiment expressed in "good." The same thinking applies to the activation dimension. This adjustment is user definable and can be set in consideration with the particular types of comment and analysis task at hand.

5.3 Simplifying Assumption of Naive Bayes Technique

A Naïve Bayes technique forms the basis for the classification scheme. The technique assumes the conditional independence of the words $\{X_i\}$ given the class $\{C_j\}$. Thus the conditional probability of a particular set of words $\{X_i\}$ occurring in a particular comment class C_j has a simpler form amenable to fast implementation via a computer algorithm. The Naïve Bayes relationship can be derived as follows

$$P(X_1, X_2, ...X_N / C_j) = P(X_1, X_2, ..., X_N, C_j) / P(C_j) \quad (11)$$

$$= P(X_1 / X_2, ..., X_N, C_j) P(X_2, X_3, .., X_N, C_J) / P(C_j) \quad (12)$$

$$= P(X_1 / X_2, .., X_N, C_j) .....P(X_n / C_j) P(C_j) / P(C_j) \quad (13)$$

Under the assumption of conditional independence of X_i given C_j

$P(X_i / X_2, ...X_N, C_j) = P(X_i / C_j)$ for all,

Moreover, the relationship simplifies to

$$P(X_1, X_2, ..., X_N / C_j) = P(X_1 / C_j) P(X_2 / C_j) ...(X_N / C_j)$$

5.3.1 Classification via Naive Bayes

By this technique, we assume that the order of words in a comment is independent of each other given the class. Thus a class is determined sorely on the specific words in a comment and their frequency of occurrence in that comment. Thus the algorithm treats the words in a comment as an unordered set or a "bag of words".

In the training phase, we classify a set of comments based on the aggregate effect of the words in that particular remark. From the training data, we estimate the vocabulary $V = \{X_1, X_2, \ldots, X_T\}$ the set of all unique words across all of the comments.
From this classification, we are able to determine the conditional probabilities of a particular word given a particular class $P(X_i / C_j)$ as

$P(X_i / C_J) = N(X_i \cap C_j) / N(C_j)$ the proportion of the number of occurrences of word X_i in class C_j across the number of words in class C_j

$P(C_j) = N(C_j) / N(C)$ the proportion of the specific class Cj across all comment classes C that occurs in the overall data set

The posterior probability $P(C_j / X_i, X_2, ..., X_N)$ which is the probability of a class given a particular set of words Xi, i = 1…N is obtained via the traditional Bayes identity

$$P(C_j / X) = P(C_j, X) / P(X) \quad (15)$$

$$= P(X / C_j) P(C_j) / P(X) \quad (16)$$

ignoring P(X) which is a common denominator, results in a measure of the maximum likelihood of a particular word set $X = \{X_1, X_2, \ldots, X_N\}$ belonging to a particular class C_j

$$P(C / X) = \max_{C_j} \{P(X / C_j) P(C_j)\} \quad (17)$$

$$= \max_{C_j} \{P(X_1 / C_j) P(X_2 / C_j) ...P(X_N / C_j) .P(C_j)\} \quad (18)$$

6. RESULTS AND PERFORMANCE

A sample of the comments analyzed to demonstrate the algorithm in this paper is found in Table 5.

Table5. Sample Comments

1.HOTLINE ISSUES - DELAY IN ANSWERING - CX SERVICE ASSISTANCE Today morning CX has called to the 444 H\L for Movie Ticket & he has waited for more than 10 mins in the line, regarding this now CX was very disappointed on our service. So pls be kind enough to chk on ths & give the call back to the CX ASAP. * Note: - Regarding this issue CX need the call back from one of our manager & CX has requested not to charge a single rupee from his no for this issue.
2.Yes,man magea prshnaya kiyapu gaman eyaa magea prshnea

wisaduwaa he's a good

3.Yes kad pin nambar signal

4.Wenath ayathana wala mema pahasukam nomati nisa

5.very good service

6.uparimaya

7.Uparima

8.think so

9.thanks

10.Super

11.Solved

12.She resolved my problem.

13.Service nallam

14.Sambanda weemata boho welawak giya nisa

15.recharge

16.Prashnayata pilithura hodin pahadili kara dima

17. Payak athulatha gataluwa nirakaranaya karanwa kiuwa. Thawamath gataluwa nirakaranaya kara natha.

18.oba ayathanaya sewawan sadaha ihala mudalak ayakarana nisa

19.no mms setting laba dunnada save kala nohaka

20.nam apahu e tika ewanna

21.Mata awashshaya u pilithurau pahadili lesa laba ganemata hakiuna.

22.mage parshnata pilithuru dunna.

23.lotari SMS stop

24.Its professional

25.ing tone sewawa ain kirima

26.I submitted Dtv reg form on 27th oct at yr crescat arcade. They told to call me on 28th wed to give the AC No

27.Hot line eka answer karapu girlge voice eka and care eka good

28.Hi kohomada? Mama mea dawas wala plan karagena yanawa mage next music video eka karanna. Song eka "Mata Rawana" :-)

29.harima pehediliwa mage getaluwa nirakaranaya kala thanks dailog

30.Good service but shortcomings due to some arrogant customer care officers

31.good men

32Good

33.getaluwa hadunagenimata noheki wiya..

34.First of all its great to be treated as a privilege customer. The reason is simple. I'm using dialog mobile connection and DTV, because dialog has the better

35.durakathanayata pilithuru denda epai eke hoda naraka kiyanna.

36.Cx need to add the CHU CHU TV which is a kids channel to the channel list.Since this channel is available on another TV connection.Cx need this channel to activate for DTV aswell.Please check on this and do the needfull. Thank you

37.Customer service personal have to be trained better cause they can't think out of the box.

38.bashawa wenaskaranna

6.1 Sentiment Rating via Enhanced Algorithm incorporating Independent Term Matching

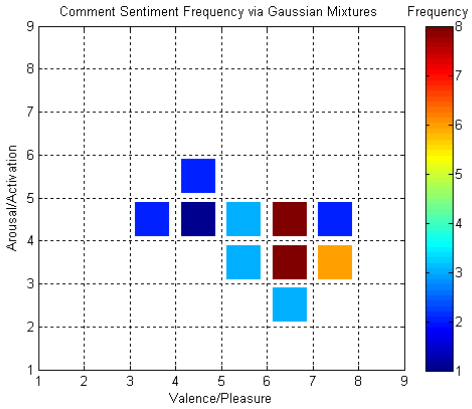


Figure 3: Overall Sentiment of the comments and their frequencies

Figure 3, depicts the distribution of the overall sentiment of each comment across all comments. It is clear the at the sentiment is positive as they occur more to the RHS of the plot while the emotion conveyed is perhaps not high in the activation dimension. A glance at the comments in Table 5, would convey such an idea when taken in a holistic sense.

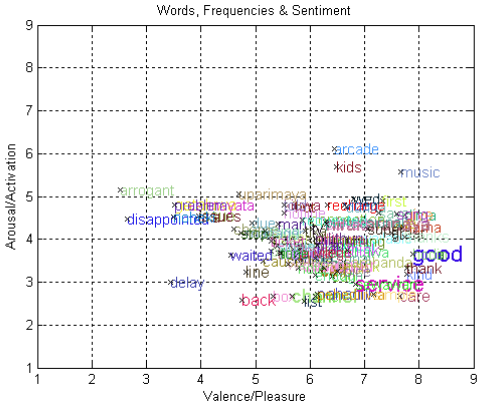


Figure 4: Sentiment of each Directory matched word and their frequency

Figure 4, Depicts the emotional content of each word appearing across the comments that have a match in the Dictionary. Words are occurring more frequently appear larger.

While in these results the sentiment was rated if there was at least one match of a comment word with the Dictionary, we should rate the sentiment if the number of matches is higher than a certain user-defined a threshold. This will further improve the accuracy of the rating process giving a better overall picture of the sentiment expressed by the customers.

6.2 Association Rule Mining words in comments

The Association Rule algorithm devised enabled rules with apriori confidence and support levels to be mined. These rules provided insight into relationships between words that are likely to occur in a comment. We cast the problem in the fashion of a market basket analysis with words as items and comments taking the place of transactions.

The rules with high support and confidence were used to obtain "keywords" that formed the antecedent and consequent terms of the selected rules. These keywords provide insight into the likelihood of a particular group of words occurring in a comment. The underlying principle of the Naïve Bayes technique is also similar where the words in a comment are considered as an unordered group, but where the word frequencies are accounted for.

Thus the rules can be used as a means to train the Naïve Bayes model with classification classes indicated by the occurrence of certain groups of keywords

6.3 Classification via Naive Bayes

Cross-validation across the training sample generated a match more than 90%.

We believe that certain comments should be excluded from the classification as they do not convey any meaningful category information that can be used to classify them. The performance will further increase with this initial preprocessing and filtering.

On predicted data, the performance was more than 75%.

6.4 Adding local language support to the Dictionary

As this use case necessitated the analysis of Sinhala and Tamil language words in addition to the English, the Dictionary was updated with such words by replicating the sentiment rating values of English words that were similar in meaning / emotional content to the non-English words and entering them against the non-English words.

6.5 Variance in max-GMM and Simple Average

Figure 5, depicts the variance between max-GMM and Simple Average methods for estimating the overall sentiment of a comment. It is seen that 90% of the time the measures are within +/- 0.5 in the case of the Valence Attribute. The CDF of the difference in the Activation attribute is tightly centred on the origin indicating hardly any variance. As discussed earlier this is also an indication that most comments convey sentiments of a single polarity and only a few comments (less than 10%) have words with conflicting emotional content.

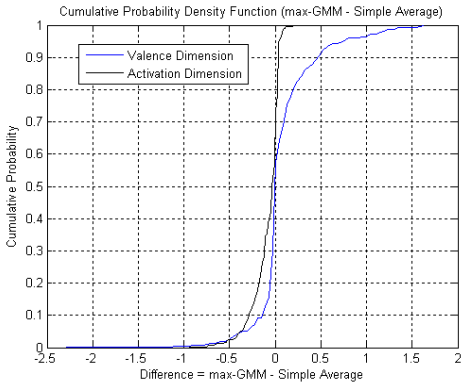


Figure 5: Variance between GMM and Simple Average measures for estimating overall comment sentiment

7. CONCLUSION AND FUTURE WORK

In this work, we demonstrated that short comments could be analyzed for their sentiment via an Independent Term Matching Scheme [3]. This technique together with its enhancements provides a way for an organization to classify the comments and also rank them by the degree of severity which other techniques using machine learning cannot achieve due to the need for a large corpus of text to train such algorithms.

Through our enhancements to the Independent Term Matching Algorithm [3], we were successfully able to rate the sentiment of comments containing negations by appropriately translating the sentiment ratings of words affected by such negations. This led to further improving the measurement accuracy of the overall emotional content of a comment.

It was also found that the sentiment ratings calculated for a typical comment in our particular application were similar whether a simple average of the sentiment ratings in each dimension was used or whether the maximum probability point of the Mixture of Gaussians was used. This is because these short comments typically have only words of a particular polarity and do not in general express more than one sentiment that's contrary to one another.

The use of Association Rule generated keywords enabled more accurate, and consistent classification of comments used to train the Naïve Bayes model resulting in improved classification / prediction performance.

Local language support in the form of an updated Dictionary enabled the algorithm to be more versatile concerning analyzing the sentiment and classifying comments containing non-English words or a mix or English and other non-English words.

We plan to improve classification performance by iteratively classifying the miss-classified comments in the prediction and adding them to the training samples. In this way we endeavour to assist the algorithm to learn from its mistakes and improve its performance with each iterative step.

8. REFERENCES

[1]. Turney, Peter (2002). "Thumbs Up or Thumbs

[2]. Down? Semantic Orientation Applied to Unsupervised Classification of Reviews".Proceedings of the Association for Computational Linguistics. pp. 417–424. arXiv:cs.LG/0212032.

[3]. Pang, Bo; Lee, Lillian; Vaithyanathan, Shivakumar (2002)."Thumbs up? Sentiment Classification using Machine Learning Techniques". Proceedings of the Conference on Empirical Methods

[4]. Ramasamy, Siddarth; "Visualization of the Sentiment of Tweets", M.Sc. Thesis, Department of Computer Science, North Carolina University, 2011

[5] web access: Sentiment_analysis https://en.wikipedia.org/wiki/Sentiment_analysis, extracted on 29 May 2016.

A Light Weight Provenance Aware Trust Negotiation Algorithm for Smart Objects in IoT

J.A.D.C. Anuradha Jayakody[1], Lakmal Rupasinghe[2], N.T Mapa[3], T.S Disanayaka[4], D.S.A.Kandawala[5], K.D.Dinusha Chathurangi[6], Krishnadeva Kesavan[7]

Department of Information Systems Engineering, Sri Lanka Institute of Information Technology (SLIIT), Malabe, Sri Lanka.

[1] j.c.jayakody@postgrad.curtin.edu.au, [2] lakmal.r@sliit.lk, [3] navoda.m@sliit.lk [4], navoda.m@sliit.lk, [5] thushari.d@sliit.lk

ABSTRACT

Internet of Things can be considered as the next big tide which advances towards the ICT realm. Many research communities have shown enthusiastic interest towards the variety of research topics which has been emerged into a discussion related to this novel concept. The research taxonomy of IoT is built upon several key pillars by considering its Complexity, Heterogeneity, and Versatility nature. Among these, security related research challenges can be considered as a key impacting domain. This particular research has been conducted with the special consideration towards Trust Negotiation among smart objects in order to satisfy provenance related criteria. Therefore this paper has suggested a light –weight, less-complex, comprehensive encryption algorithm by applying shuffling techniques in order to satisfy the origin identification.

***Keywords*— Internet of Things (IoT), Security in IOT, Provenance, Trust Negotiation, Light – Weight Encryption**

1. INTRODUCTION

Current behavioral patterns depict that people around the world are more encourage towards consuming the services provided by the Internet to accomplish their day to day wants, needs and tasks. However, the resent observation represents an unprecedented consuming pattern with regards to the consumption of applications and services provided by the Internet. Based on the current context it is predictable that within the next years or decades to come the need of Internet-based services are going to be exceptionally high, and more people will be in the urge to access the global information contents. In such perspective, conventional methods will not be able to produce the expected outcomes that are been required. Therefore this eager to extend the interconnections has paved new ways to establish a big leap in the future of Internet. As a result, the concept of IOT has been emerging into the discussion.

IOT can be defined as an Umbrella concept which comprises all these desired aspects based on the paradigm of computing and communication. It has been built upon the idea, where the notion of interconnected smart devices acquire the main key objectives of anywhere - anytime - anything connectivity [2].

Since IOT can be defined as a futuristic technology trend, it consists of several challenging research domains [3].

Based on the considerations above the main scope of this investigation has been focused towards the role of the Security Domain in IoT. Since security, considerations has been a critical component it has been span across by reference to several research areas related to IoT. The following diagram depicts the taxonomy of the most demanded research areas relevant to the Security Domain of IoT.

Fig.1: Security in IoT

By reference to the context mentioned above, the core research focus of this paper has been narrowed down to the Trust based challenge objectives related to IoT. The follow- up content will clearly

clarify the actual need of embedding Trust into the IoT Infrastructure.

Internet of Things has been able to compose the technical and research savvy communities into diverse directions. The overall context of IoT comprises of an inter-connected, inter- related, the ever-growing system of physical objects, smart devices - embedded with electronics, software, and sensors.

The network connectivity which binds the entire IoT infrastructure thrives on a timely evolving spectrum of data collections generated by all most any object connected to the IoT network. These sets of data collections can be directed towards several stages of processing and ultimately can be manipulated in different ways by the application of particular logics. Ultimately the resulting Information Workflow can be used as the reference base in order to perform series of decision making related activities. Since the most vital attention has been focused towards the fetched data which is then directed towards into different levels of processing, ensuring the trustworthiness and the original ownership of the individual objects which produces the data has been a vital challenge. This particular objective has initiated the necessity of confirming certain level of data trust which can be reached by the application of the concepts related to Data Provenance. In order to ensure the provenance of an object, the object wise information extraction should be done at the initial point where the node was introduced to the IoT network base.

Based on the information above the base, the main objective of this research is to deliver a less-complex, light- weight trust negotiation algorithm to satisfy the requirement of provenance negotiation among the heterogeneous modes in IoT.

The remainder of the paper is organized as follows. In Section II, the paper has introduced and compared the different visions of the solutions available to the related subject, which is available from the literature. The system overview and the descriptive information about the proposed trust negotiation algorithm is presented subsequently in Section III and Section.

IV. The Final content Section V, concludes the paper and presents further extensions that could be performed which is aligned to the subject.

2. LITERATURE SURVEY

One of the most popular topics nowadays is Internet of Things (IoT) causes a high impact on several aspects of everyday life and behavior of potential users. Main advantage of the IOT is that it provides development of a huge number of applications in various domains such as

- Transportation and logistics domain.
- Healthcare domain.
- Smart environment (home, office, plant) domain.
- Personal and social domain [1].

Within this huge number of applications a wide range of individual objects identifiable via own IP addresses or similar identifier. Data provenance identifies the origin of data and processes operations on it will help to assure security requirements such as Integrity and Confidentiality of collecting data in the IoT. To get accurate trust level, several requirements need to be satisfied such as [2]:

- Completeness of Information
- Integrity
- Availability
- Confidentiality
- Efficiency
- Privacy
- Trust

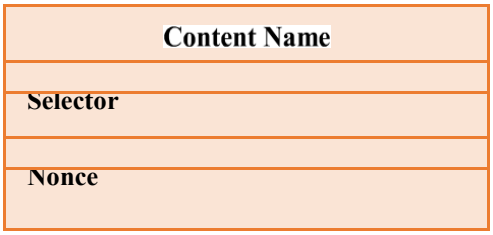
When it comes to the factor “Trust” it has a large number of Definitions and widely used definition is the one provided by Blaze and Feigenbaum, which refers to security policies regulating accesses to resources and credentials that are required to satisfy defined policies. At present a limited number of solutions available related to the identity management and access control issues. Most popular approaches include keynote and trust builder. However, any of these do not lend them to a straightforward application to the IOT domain due to high computational requirements that they impose [3]. In the paper of Javier Suarez, Jose Quevedo, Ivan Vidal, Daniel Corujo, Jaime Garcia-

Reinosa, Rui L. Aguiar on "A Secure IoT management architecture based on Information-Centric Networking" they had proposed a new architecture to improve trust between objects. According to them, it was a gargantuan challenge to develop an architecture which met all the requirements mentioned below.

- Scalability
- Energy Efficiency
- Self-organization
- Semantic Interoperability
- Privacy
- Security
- Computational ability of devices

Their architecture provides a generic and flexible platform that allows the appropriate operation of IoT devices with in a delimited ICN network domain, such as an organization or residential environment. Their design was designed to support a wide range of devices of different types and capabilities. The main component of their architecture is the gateway which acts as an intermediary communication element between clients and IoT devices. When it comes to the topic of discovery and registration of devices gateway, manually authenticate and configure the security mechanisms. In Object authentication procedure they have calculated a cryptographic digest of its public key, and a key locator, indicating where that public key can be obtained. The digital signature included by the gate way in data packets.

Interest Packet



Data packet

Fig.2: Data Packet and Interest packet

The encryption with the public key of the client used for ensuring a client is a legitimate object. When a client receives a data packet, client validates the digital signature and authenticate the gateway as the originator of the data packet and verify its Integrity [4].With the use of encryption mechanisms they have achieved the trust between objects

3. SYSTEM OVERVIEW

The concept of trust has divergent interpretations based on different contexts. Variety of definitions has been delivered on the adapted perspective. In the notion of IOT, still, there is no consensus definition related to the concept of trust even though its importance has been highly recognized. Since the peer – to – peer negotiation of different smart objects inside an IOT network is the key element to be established prior to establishing a communication session, ensuring it has been based on sensitive trust negotiation is a key point. This requirement leads to maintaining a mutual trust between the peer nodes of the IOT network. In order to perform such operation provenance, related information should be maintained in a proper way. The prevalent mechanism to perform such operation is by including the provenance information as an additional attribute to the routing information. Since provenance is concerned towards verifying the origin-related information of a smart node, the sensitivity of that information must be preserved. Therefore as means of persisting the provenance, this paper has suggested a Trust Negotiation mechanism focusing on Integrity constraint, by which the shared provenance information is protected from intruders. The structure of the proposed algorithm is described in the follow-up content.

Nodes which are willing to take part in the IOT network are configured with a key that is agreed upon the nodes. This pre-shared key is only known to the peer nodes. When the nodes need to exchange information via messages, the message proceeds through several steps. First of all the original message is taken. Then a random portion is selected out from the original message. The random selection depends on several characteristics of the same original message. Next, the message will be encrypted using a lightweight encryption algorithm by the use of shuffling mechanism and based on the set of properties related to the pre-shared key.

Finally, the shuffling algorithm produces the cipher text (Encrypted text) which is encapsulated in the message to be sent to the other side.

The receiving node takes the message and extract the cipher message portion and decrypt it with the pre-shared key in its possession. The integrity of the message is ensured if the received message portion and the decrypted message are matching with each other.

Comprehensive description of the aforementioned light- weighted encryption algorithm is presented in the next section.

4. SHUFFLING ALGORITHM

To apply shuffling algorithm following parameters need to be identified.

- Number of bits in the key
- No of zeros in the key bit stream
- No of ones in the key bit stream
- No of bytes in the message
- No of zeros in the randomly selected byte stream
- No of ones in the randomly selected byte stream

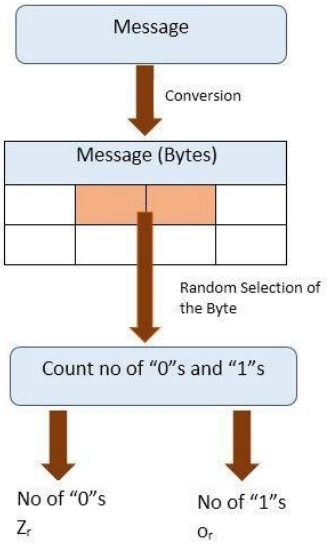


Fig.3: Shuffling algorithm's flow of counting 1's and 0's of message

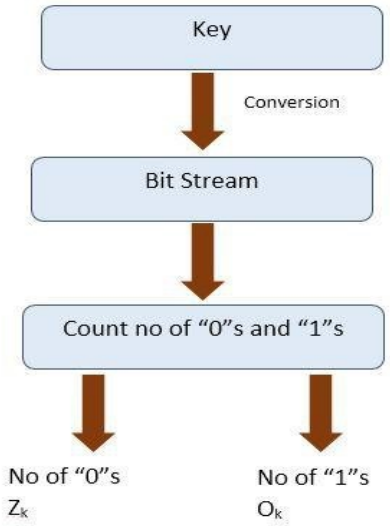


Fig.4: Shuffling algorithm's flow of counting 1's and 0's of key

When it comes to the random selection of the byte, the position of the byte is calculated using the following equation. In that case, we calculate the no of bytes in the message and divide it from the maximum no of zeros and number of ones.

RB - Random selection of the byte

N - No of bytes in the message

M - Max (number of ones, number of zeros)

$$RB = N / M \rightarrow (1)$$

Then we combine the no of zeros in randomly selected byte and key bit stream into a single parameter called " Z "and no of ones into a parameter called "O" using following equations.

Z - Combined number of zeros

Zr - Number of zeros in the key

Zk - Number of zeros in the random byte

$$Z = Zr + Zk \rightarrow (2)$$

O - Combined number of zeros

Or - Number of zeros in the key

Ok - Number of zeros in the random byte

$$O = Or + Ok \quad (3)$$

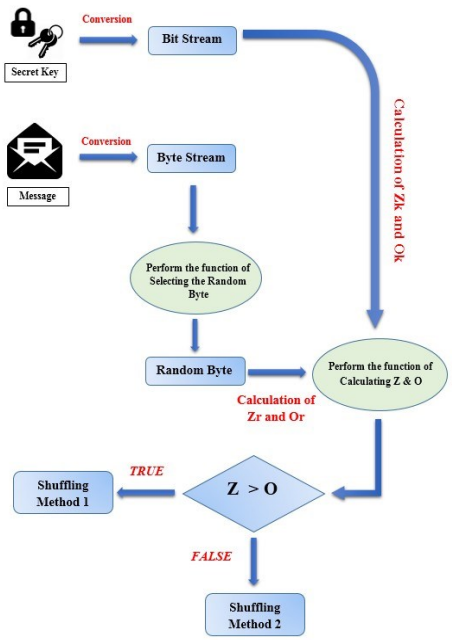


Fig.5: Flow of Shuffling Algorithm

The shuffling method will be selected after performing the function of calculating Z and O. If Z>O then the shuffling method 1 will be used. Else method 2 will be used.

A. Shuffling Process: Method 1

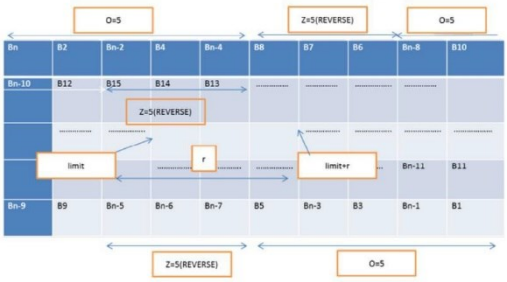


Fig.6: Example of Method 1

Here for "O" number of bytes we interchange the bytes with the last "O" number of bytes. We take "Z" number of bytes after the "O" number of bytes and reverse them in both front and the back of the message. Likewise, we continue the process until we meet a limit which is denoted by the variable "limit". Here we assume that "Z" is equal to 3 and "O" is equal to 5.

A. Shuffling Process: Method 2

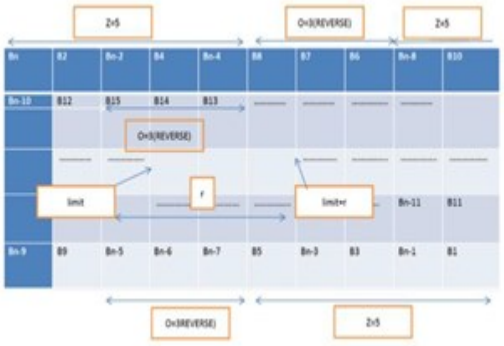


Fig.7: Example of Method 2

Here for "Z" number of bytes we interchange the bytes with last corresponding "Z" number of bytes. We take "O" number of bytes after the "Z" number of bytes and reverse them in both front ends the back ends of the message. Likewise, we repeat the process until we meet the variable "limit." Here we assume that "Z" is equal to 5 and "O" is equal to 3.

The "Limit" variable will be calculated using the following equation.

$$Limit = (tb / (O+Z)*2))*(O+Z) \longrightarrow (4)$$

Above given equation (4) will be calculated based on integer division. The parameter "tb" is calculated relative to the length of the message (tb = message.length ()).

5. CONCLUSION AND FUTURE WORK

Nodes which exists in an IoT shows versatile characteristics due to its heterogeneous nature. This requirement has led to ensuring a certain level of origin based trust related to the associated network. Trust Negotiation among the smart nodes has always been a challenging goal since the devices present are equipped with different levels of power and computational capabilities. Based on these key factors this research paper has presented a light-weight, less- complex trust negotiation algorithm with the application of simple shuffling mechanisms. When the focus is extended towards the future, the structure of the algorithm can be further optimized to achieve less computational and fewer power consumptions patterns.

6. ACKNOWLEDGEMENTS

This work has been supported by Sri Lanka Institute of Information Technology, Malabe, Sri Lanka.

7. REFERENCES

[1] Luigi Atzori a, Antonio Iera b, Giacomo Morabito c, "The Internet of Things: A survey", DIEE, University of Cagliari, Italy, University ''Mediterranea" of Reggio Calabria, Italy, University of Catania, Italy, 2012.

[2] Sabine Bauer, "Data Provenance in the Internet of Things", IT-Security Group, University Passau, 2013.

[3] Daniele Miorandi , Sabrina Sicari , Francesco De Pellegrini , Imrich Chlamtac "Internet of things: Vision, applications and research challenges", in International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery, Int. Conf., 2013.

[4] Javier Suarez,Jose Quevedo,Ivan Vidal,Daniel Corujo,Jaime Garcia- Reinoso,Rui L.Aguiar., "A Secure IoT management architecture based on Information –Centric Networking", University of Journal of Network and Computer Applications, 2016.

[5] John A. Stankovic," Research Directions for the Internet of Things"Computer. Sci. Dept., Univ. of Virginia, Charlottesville, VA, USA,2014

[6] "The Internet of Things: An Overview", 2015. [Online].Available:https://www.internetsociety.org/sites/default/files/ISOC-IoT-Overview-20151014_0.pdf

[7] Friedemann Mattern and Christian Floerkemeier," From the Internet of Computers to the Internet of Things",2009

[8] "Internet of Things", 2015.[Online].Available:http://www.cisco.com/c/en/us/solutions/internet-of-things/overview.html

[9] Ala Al-Fuqaha et-al," Internet of Things: A Survey on Enabling Technologies, Protocols, and Applications", 2015

[10] "Security In The Internet Ofthings",2015.[Online].Available:http://www.windriver.com/whitep apers/security-in-the-internet-of-things/wr_security-in-the-internet-of- things.pdf

[11] L. Atzori, A. Iera, and G. Morabito, "The Internet of Things: A survey," Comput. Netw. vol. 54, no. 15, pp. 2787–2805, Oct. 2010.

[12] R. Khan, S. U. Khan, R. Zaheer, and S. Khan, "Future Internet: The Internet of Things architecture, possible applications and key challenges," in Proc. 10th Int. Conf. FIT, 2012, pp. 257–260.

[13] J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami, "Internet of Things (IoT): vision, architectural elements, and future directions," Future Gener. Comput. Syst., vol. 29, no. 7, pp. 1645–1660, Sep. 2013.

[14] P. Lopez, D. Fernandez, A. J. Jara, and A. F. Skarmeta, "Survey of Internet of Things technologies for clinical environments," in Proc. 27th Int. Conf. WAINA, 2013, pp. 1349–1354.

[15] D. Yang, F. Liu, and Y. Liang, "A survey of the Internet of Things," in Proc. 1st ICEBI, 2010, pp. 358–366.

[16] A. Gluhak et al., "A survey on facilities for experimental Internet of Things research," IEEE Commun. Mag., vol. 49, no. 11, pp. 58–67, Nov. 2011.

[17] Z. Sheng et al., "A survey on the IETF protocol suite for the Internet of Things: Standards, challenges, and opportunities," IEEE Wireless Commun., vol. 20, no. 6, pp. 91–98, Dec. 2013.

[18] J. Gantz and D. Reinsel, "The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the Far East," IDC iView: IDC Anal. Future, vol. 2007, pp. 1–16, Dec. 2012.

[19] M. Wu, T. J. Lu, F. Y. Ling, J. Sun, and H. Y. Du, "Research on the architecture of Internet of Things," in Proc. 3rd ICACTE, 2010, pp. V5- 484–V5-487.

[20] EU FP7 Internet of Things Architecture Project, Sep. 18, 2014. [Online]. Available: http://www.iot-a.eu/public

The Simulation – Based Solution to Detection by Collectively Decomposing Factors of IMU and Image Process Data in the Indoor Environment

J.A.D.C.Anuradha Jayakody [1],R.G.P Dulshan[2], D.C Meegahawatta[3] , H.M.K.S.B Herath[4], L.P.A.D Pathirathne[5]

Department of Information Technology, Sri Lanka Institute of Information Technology (SLIIT), Malabe, Sri Lanka.

[1]anuradha.j@sliit.lk, [2]dulshanqw@gmail.com, [3]chamikaradinuka@gmail.com, [4]kalanaherath15@gmail.com, [5]ashan.dileepa@gmail.com

ABSTRACT

Existing indoor navigation system face with many different technical and usability problems because of the localization. In this paper presents indoor navigation simulation-based solution that can apply to the real implementation. The proposed system has several benefits and has the potential to increase the usability of the scheme.

Navigation is the process of monitoring and controlling the movement of and item from an origin to a destination along the path. Navigation system provides reading monitoring and updating the movement of one's position and guiding by intelligible visuals. Audible or Tangible means while she is traveling on an intended route.

In this simulation movement of the user is shown by the navigation and based on the simulation human movement can identify the path, location, the remaining distance to their final destination. Moreover, this will be critical for the user to determine their exact path. Any number of user can use this at the same time, and it has the facility to simulate the crowd sourced.

Environment. Further, this research work focuses database optimization on reducing access time as an optimized solution.

Based on the simulated results authors plan to develop an algorithm that facilitates to indoor navigation with localized information and plan to discuss test results with evaluation.

Keywords- **IMU, REST, ath_H_ID, ImgD , Labelname(LN),PlaceDescription(PD),StartingX (SX),EndingX(EX),StartingY(SY),EndingY(EY) ,FloorID(FID),BuildingID(BID).**

I. INTRODUCTION

This context is providing a complete idea about Indoor navigation and database optimization mechanisms. This research can be described three main components. Those are,

- Data Collection
- Data storing
- Utilizing stored data

As an initial step for the research, it needs a huge data collection first. That data collection will be the input for the database optimizing part. Therefore the research team has populated this huge data collection by using two main techniques which commonly available in current indoor navigation systems.

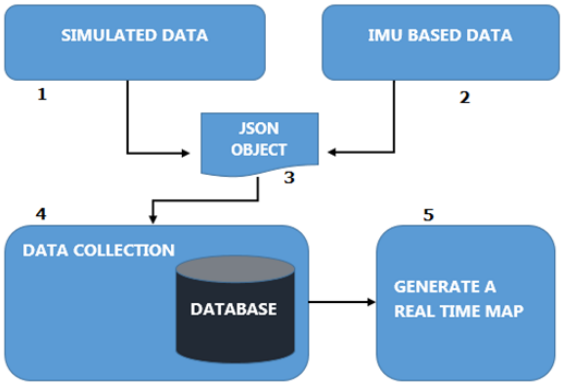


Figure 1: Overall System Architecture

1. Simulated Data

Under this component, it describes a simulated indoor navigation model which generates hundreds of records within few seconds. It will produce live data collection base on the user's requirement. What the user has to do is only select the number of user and their destinations. After that, the simulator navigates selected user to their destination, and meanwhile, it collects all the information which belong to user's each and every movement.

Ex: User current x coordinates, User current x coordinates, R-value, Theta Value, Obstacle detection, environment details .(door info, wall info)

2. IMU-Based data

Under this component, it describes real-time data. Here a real-time navigation based data collection will be generated on behalf of simulated data collection. It uses IMU (inertial measurement unit) techniques to generate those real-time data collection. What the user has to do is only walking along the path with his smart phone. Smart phone itself identify the user's current coordinates and step counts. Base on IMU data it produces same data collection as simulated data collection.
Ex: User current x coordinates, User current x coordinates, R-value, Theta Value, Obstacle detection, environment details. (door info, wall info using image processing)

3. JSON Object

In this research, it generates huge data collection by two ways. That data should be accessed the Database at the same time without any delay. Here a JSON object is used in order to make the process speed up and real time.

Why JSON?

- JSON is a much simpler.
- Smaller message size
- More structural information in the document
- The speed of processing.
- JSON is easier to read.
- JSON requires fewer tags

In every database call, all the values will be converted to a JSON array and send them to the database through the web service.

4. Data Collection

In this component, it will optimize and store all the data comes from the user. (May be the simulator or a real user). For the database optimization part, it has been used several techniques.
Ex:

- Indexing
- Memcaching
- Profiling
- Costing
- Stored procedures
- Generate a real time map

Up to this level, we have collected a huge data set, and that data set has been optimized by using several optimizing mechanisms. Therefor at this step, it is going to be utilized that optimized data collection by generating a real time map. Real time in the sense, it can automatically update the environmental changes time to time.
Ex: Door information, wall information, and moveable object information, etc.

II COLLECT DATA FROM SIMULATOR

In the simulator, user can select the number of users and their starting point and destinations respectively. Then once the app user started the navigation process, it will collect each and every movement information of every user. Following variables will be captured during the navigation process.

1. User's current x coordinates.
2. User's current y coordinates.
3. R-value
4. Theta value
5. Is obstacles ahead?
6. Obstacle details.
7. Environment details.

1. User's current X and Y coordinate

The map of the simulator will hold the following ratio with the real world.

Map ratio : 8 pixel = 0.8 m

Therefor every time a marker passes 8 pixels, it will be counted as 1 step. Step count will be captured according to that scenario.

2. User is angle. (Theta value)

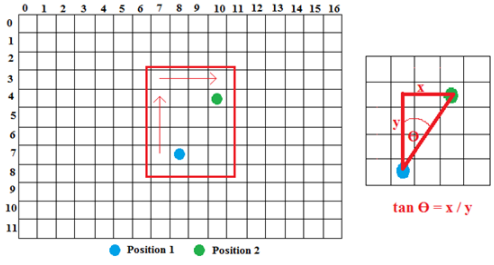


Figure 2 : User's Angel

Let's say the user moved position 1 to position 2 as mentioned in the above picture. Then the Theta value will be calculated as follows.

Calculation:

$$\text{Tan } \Theta = (10 - 7) / (8 - 4)$$
$$= \tan^{-1} [0.75]$$
$$= 36.869^{0}$$
$$= 37^{0}$$

3. *Obstacle detection*

Normally once a user starts his or her navigation it keeps an interested area. It means, the user well aware about all the movable and non-movable objects with in that area as mentioned in the following image.

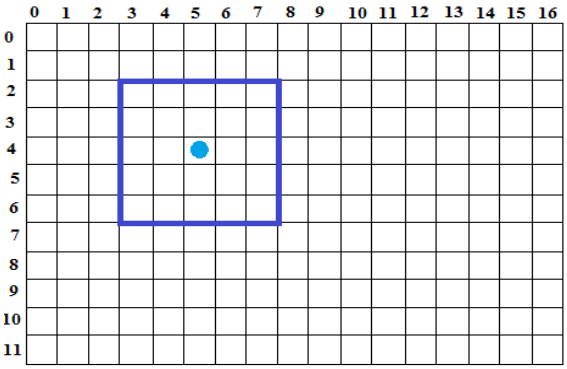


Figure 3 : Obstacle Detection

Once the user detects some object within his or her interested area, it identifies whether it is a movable object or non-movable object.

If it is a movable object, the user considers it as obstacles and calculates the distance and angle of the obstacles based on user's current position. At the same time, that information will be sent to the database as well.

If it is a non-movable object user identifies it is a door object. There can be multiple objects within the user's interested area. Still, it identifies each object and updates their status. (Status = whether the door is open or not)

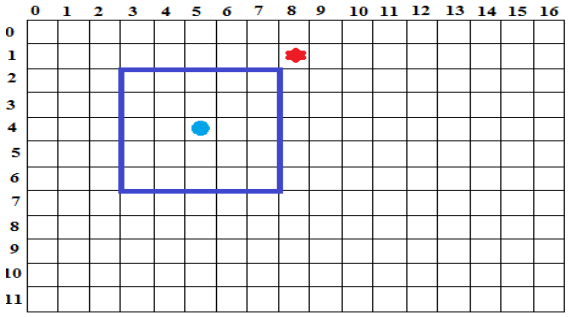


Figure 4 : Obstacle Identification Non Moveable

4. *Data communication between simulator and the server*

The communication between the server and the Navigation system is done using the JSON (JavaScript Object Notation). It is a lightweight data interchange format. It is based on a subset of the JavaScript Programming Language. To construct

HTTP requests and responses plain text or JavaScript Object Notation (JSON) is used. However, to send multicast messages must use JSON. Multicast messages mean the ability to send the same message to multiple devices simultaneously.

III. MAP GENERATION

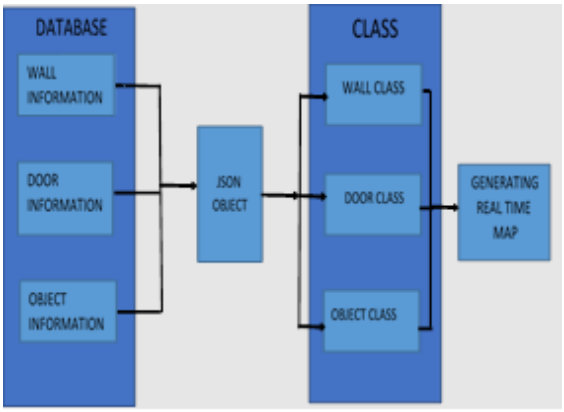


Figure 5 : Map generation overview

In the above diagram is explained the how the data is taken from the database. In the database, there is three tables called wall info, door info, object info, and we collect that data through the JSON object and build the three class call wall class, door class, object class. So final using that data and generate the map. All the object is shown on the map according to the given data.

This section introduces the methods and techniques that we have planned to follow in order to achieve our goal in this research project.

While navigation systems for outdoor environments are already available, navigation within buildings still poses a challenge. It's difficult to create a map in accordance with navigation. This map is created using the wall information of x, y coordinates. Below mention is some steps in accordance.

- Gathering data from the database.
- Identify the wall information related to x and y coordinates.
- This identifies the situation related to physical.(doors with x, y coordinates as follows if a door is closed it represented by red if a door is closed it represented by green)
- When the users move, the object in the building is automatically identified and plotted on the map.

When the map is created, these four fact is mainly taken into consideration and time to time the database access to create a map. The outcome is the real time map. The system overview is mainly based on client-server architecture. The map and the navigation pointer will be displayed on the client side with the smart device. All the sensor data will be collected in the actual environment by the smart device and generate the routing data from the virtual environment in a desktop application. The desktop application and mobile device are exchanging data through a communication protocol. The sensor information will be partially processed by the application and send them to the server to build the map.

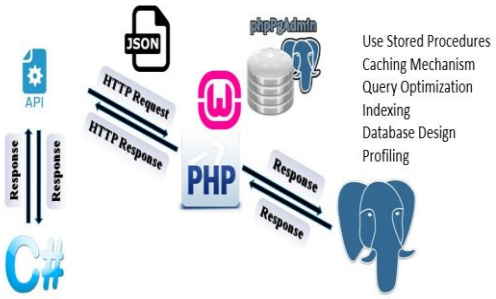


Figure 6 : Overview of the database optimization

IV. DATABASE OPTIMIZATION

Figure 6 diagram is explained that the overall view of how the database is optimized with the help of the API created by using c# dll, web service written by using PHP and PostgreSQL database management system. Several Technics such as stored procedures, caching mechanism, changing the database design, indexing, query rewriting, functions and profiling are used to optimize data inside database management System.

According to Figure 6 diagram, C# application (simulator) and PostgreSQL database are there. Web service based REST architecture is used to transfer data between the C# application and the PostgreSQL database. REST is web standards based architecture and uses HTTP Protocol for data communication. It revolves around resource where every component is a resource, and a resource is accessed by a common interface using HTTP standard methods. REST is almost always going to be faster and leading to lower bandwidth.

API is created using C# dll to interact C# application and web service. This dll consists of the database connector and the API functions (insert, update, and delete data within the table) to send and received required information. That data should be accessed the Database at the same time without any delay. For that purpose a JSON object is used in order to make the process speed up and real time. Main queries are executed inside postgrSQL database management system by using stored procedures so that it can be shared by a number of programs.

Reading data from a disk is still much slower than reading data from memory. In order to maintain an overall speed of the real-time database, memory is used to cache the most recent read data so next time same data is requested, the system will read directly from memory instead of disk.

The architectural overview (Figure.7) shows how the database schema organizes tables to store indoor data. In the initial stage, it includes features such as database connections, function call, and function return. Those function should occur respectively to fulfill the manipulation of the relation.

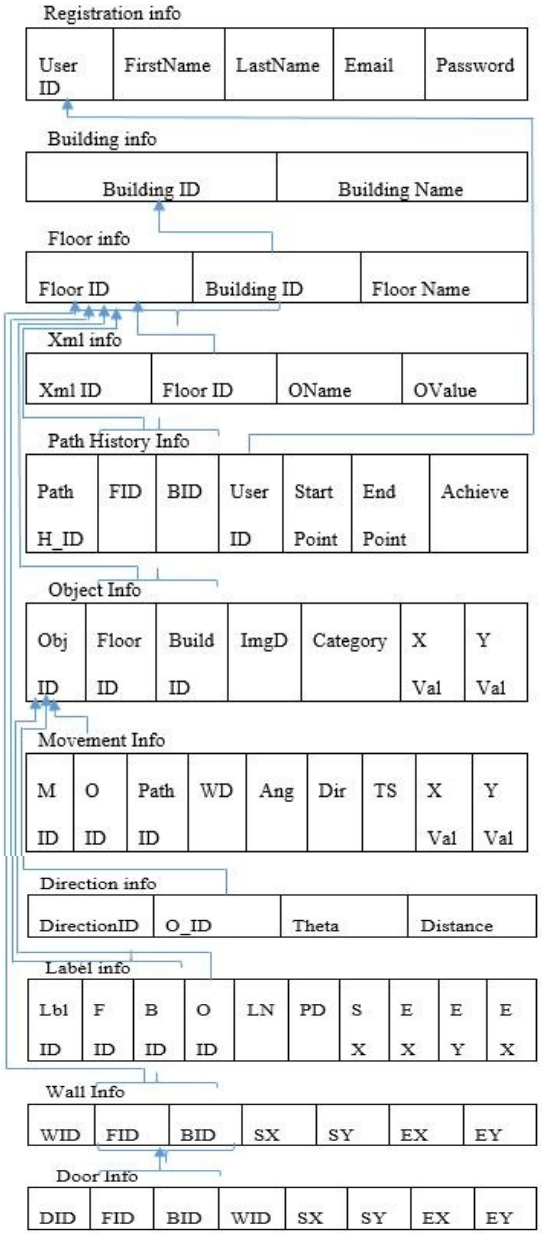


Figure 7: Design to present an organization of indoor navigation data

The proposed schema (Figure.7) clearly shows the table relationships and the primary and foreign keys to denote the information of the relational tables that can be used to develop an indoor map that can be support the people who are keenly waiting for using it in the indoor premises. The "object info" table has "Object id" as a master key which is generated automatically by the schema. It follows that object description field. An object id should refer to the "Label info" table attribute "Object id" as a foreign key and then get the received label information. The "Movement info" table should need to access the "Path history info" table to get the routing details. The time stamp attribute should refer to the "Movement info" table and "Path history id" as a foreign key. An "Object id" of the "Object Info" table should refer to the "Direction info" table to find the direction details through "Object id" of "Direction Info" table. "Floor info" has "floor id" and "building id" as a composite master key. That master key goes as a foreign key to "building info". Every floor has a specific xml file to map details. "Wall info" has "wall id," "floor id" and "building id" as a composite master key and that master key goes as a foreign key to "Door Info" table.

V. CONCLUSIONS & FUTURE

There is a great number of scopes for future improvements to our current implementation that can enhance both the performance and the user experience of the application.

ACKNOWLEDGEMENT

This work has been supported by Sri Lanka Institute of Information Technology, Malabe, Sri Lanka.

REFERENCES

[1] Y. Alon, A. Ferencz, and A. Shashua. Off-road path following using region classification and geometric projection constraints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2006*, volume 1, pages 689–696, June 2006.

[2] Levente Bagi. Pedometer - Android app.

[3] P. Bahl and Padmanabhan V.N. Radar: an in-building rf-based user location and tracking system. In 2, pages 775 – 784, March 2000.

[4] Sudarshan S. Chawath. Marker-based localizing for indoor navigation. In Intelligent Transportation Systems Conference, 2007. ITSC 2007. IEEE, pages 885 – 890, September 2007.

[5] Jaewoo Chung, Matt Donahoe, Chris Schmandt, and et al. Indoor location sensing using geo-magnetism mobisys, 2011.

[6] Sinan Gezici, Tian Zhi, G.B. Giannakis, and et al. Localization via ultra-wideband radios: a look at positioning aspects for future sensor networks. Signal Processing Magazine, IEEE, 22:70–84, July 2005

[7] Google. Android sensor guide.

[8] Kamol Kaemarungsi and Krishnamurthy Prashant. Modeling of indoor positioning systems based on location fingerprinting, 2004.

[9] L. Klingbeil and T. Wark. A wireless sensor network for real-time indoor localization and motion monitoring. In International Conference on Information Processing in Sensor Networks, 2008., pages 39–50, 2008.

[10] Jo Agila Bitsch Link, Felix Gerdsmeier, Paul Smith, and Klaus Wehrle. Indoor navigation on wheels (and on foot) using smartphones. In Proceedings of the 2012 International Conference on Indoor Positioning and Indoor Navigation (IPIN), November 2012.

[11] Android market. Runtastic app.

[12] Oliver Woodman and Robert Harle. Pedestrian localization for indoor environments. In Proceedings of the Tenth International Conference on Ubiquitous Computing (UbiComp 08), pages 362–368, 2008.

[13] Fan Li, Chunshui Zhao, Guanzhong Ding, and et al. A reliable and accurate indoor localization method using phone inertial sensors. In Proceedings of the 2012 ACM Conference on Ubiquitous Computing, pages 421–430, September 2012.

[14] Alessandro Mulloni, Daniel Wagner, Dieter Schmalstieg, and Istvan Barakonyi. Indoor positioning and navigation with camera phones. In *Pervasive Computing, IEEE*, volume 8, pages 22–31, April 2009. 7

[15] MySQL. [Online] Available: https://www.mysql.com [Feb, 1, 2016].

[16] PostgreSQL. [Online] Available: http://www.postgresql.org. Web. 1 Feb. 2016.

[17] SQL Server 2014. [Online] Available: https://www.microsoft.com/en-us/server-cloud/products/sql-server. 3 Feb. 2016.

[18] Oracle. [Online] Available: http://www.oracle.com/index.html. [Feb, 2, 2016].

Predictive Analytics with online data for WSO2 Machine Learner with the support of Ensemble method

Lakmal Rupasinghe[1], Heshani Herath[2], Ishani Pathinayake[3], Ashani Diaz[4],Indujayani Karthigesu[5]
Sri Lanka Institute of Information Technology, Sri Lanka
lakmalr@gmail.com, phone +94-77-756 1061[1]
heshani7.herath@gmail.com, phone +94-71-739 0012[2]
ishanipathinayake@gmail.com, phone +94-71-074 9024[3]

ABSTRACT
Different types of malware prevail in a world of innumerable computer users who constantly struggle with threats from malware sources such as the internet, local networks, and portable drives and so on. Security specialists and malware developers have been at a tug of war with each other as security specialists try to use all possible techniques to overcome the problems arising from malicious software while malware developers try to bypass these implemented security measures. It can be observed from records that each day, nearly 1 million new malware threats are released to the world. Therefore it is evident that there is an indispensable requirement of a proper malware identification mechanism. Typically, antivirus software is based on a signature definition system which keeps updating from the internet and thus keeping track of known malware. While this was sufficient sometime back, it does not cater to the current requirement of identifying malware. Due to the advancement in technology, malware developers have been able to create malware that is able to obfuscate themselves thus giving birth to polymorphic malware. In this study, we closely observed the behavior of malware, tried to understand how they work, their different types, dissemination of malware and detecting/defending mechanism in order to contribute to the process of security enhancement and came up with the solution of integrating Machine Learning to our current problem.

***Keywords*: Malware family classification, predictive analytics, HTTP Streaming data**

I. INTRODUCTION

The Malware is a topic widely spoken in the world of information technology. Malware, being the short term for malicious software consists of code snippets, scripts, active content etc. The term has invaded the digital world in such a manner that everybody knows its meaning and has taken steps to prevent their computers from being infected by them as they cause severe problems leading to the loss of privacy, increase unauthorized access to system resources and other abusive behavior and even access to the system unstable. Malware can reach the systems in different ways and through multiple media. Reports show that the most common entrance of malware is through the downloading process via the internet. Malware acts in many different ways. There is some malware that will not totally harm the system but will affect the performance and create overload process; certain other malware can act as spyware in which case the malware will hide in the system, and the average anti-virus software will not be able to detect them. This hidden malware will send critical information about the computer to the source. By looking at the above challenges, it is imperative to carry out a thorough investigation to understand the behavior of malware for better detection.

II. OBJECTIVES

There have been some efforts to use few machine learning and data mining for the purpose of identifying new or unknown malicious code. However, we are mainly focused on the identification of malicious code in HTTP data stream. Malicious software in its various forms constitutes a serious threat to Internet security threat. Therefore, this area has received much consideration in the research community, and many different concepts and techniques for the analysis of malware have been proposed. Our goal is, provided that the system predicts an URL to be malicious, and classifying the malware families.

In this paper, our goal is to get the lowest number of false positives as possible, by using a simple combination of various stages of the different versions of the simple algorithm. Other automated classification algorithms could also be used. We are going to extract data from HTTP streams and going to create feature table with abnormal parameters with the use of extracted data.

A set of characteristics is calculated for the HTTP data streams in training or testing datasets on the basis of many possible ways of analysing malware. Then classifying the malware using machine learning algorithm. We used three datasets a training dataset, a test dataset. The training and test data must be

representative of the web servers to be protected, and the attacks used for testing need to illustrate the range of attacks existing today.

III. SYSTEM OVERVIEW

This section describes the system architecture and the design of the proposed system. It mainly focused on few specific areas. HTTP stream (Online data stream) can contain millions of malware and important data for the day today usage as well. The proposed system is mainly focusing on capture those data which comes along with an HTTP stream. The system will identify parameters which come along with the HTTP data stream. Those parameters can be changed depend on the stream. Following we have listed some common parameters which come along with HTTP stream.

TABLE I. PARAMETERS OF HTTP DATA STREAM

Buffer size	The amount of memory allocated for sending content
Client port	A range of client ports for communication.
version	A string specifying the version of HTTP to use
Host name	The host name or IP number for the machine.
port	The socket port number.
Proxy host	The host name of the proxy server
Proxy port	The port number of the proxy server
Stream type	Whether the stream is a text stream, image stream, video, etc.

The proposed system will identify parameters of the HTTP data stream and recognize the variations of each parameter. If there is any unusual behavior system will notify that. Since malware have polymorphic behavior, there is a high possibility even virus guards will not detect the malware. A computer malware is a self-replicating computer program that operates without the consent of the user. It spreads by attaching a copy of itself to some part of a program file, such as a spread sheet or word processor. Malware also attack boot records and master boot records, which contain the information a computer uses to start up [1]. A polymorphic malware creates an infection in a computer which is known as a polymorphic infection that creates copies of itself, each with different copy to fool a malware detection and users. Variations are usually various forms of encryption or other signatures to make it more difficult for a malware detection program to find and remove the malware from a computer [2]. Sometimes those polymorphic malware differ one from the other with only one byte of change. Existing virus guards detect this change, and it checks the malicious software with their database which contains malware patterns. If they cannot find the new behavior in the database, it identifies it as a new malware and add it to their database. Later they give it as an update. However, if we consider what happens there, the pervious malware changes one byte of it and act like a new malware. However, it is not a new malware. It is a member of previous malware's family. If we can classify those malware into families, it would be a more effective than a virus guard. For this purpose only we use machine learning.

For malware family classification the proposed system uses Naïve Bayes algorithm [3-8], Support Vector Machine algorithm (SVM) [9-16] and Random Forest classification algorithm [17-19]. These machine learning algorithms support for real-time data classifications. Since the system gets three outputs from above three algorithms, the proposed system implements an ensemble method which gives the most accurate prediction.

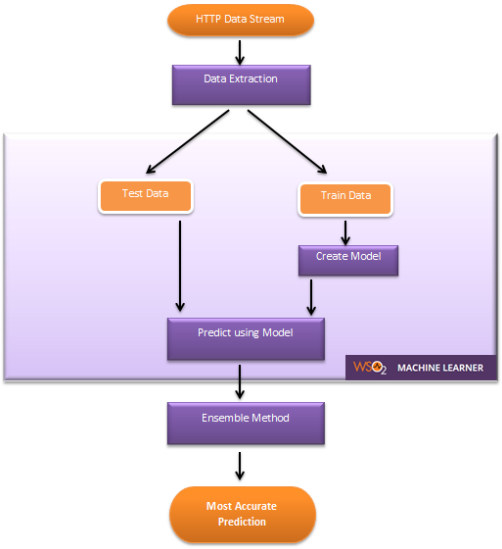


FIGURE II. ARCHITECTURE SYSTEM

IV. METHODOLOGY

This section includes detailed descriptions of the techniques and mechanism employed to make this project a reality. The descriptions include how software implementation of our project is carried out, what are the materials and data needed, and how they will be collected. It also includes time frames and schedules that are required in achieving its objectives. In addition to them, the research areas that we have identified in order to carry out this project are explained rationally.

Creating another machine learner is not the goal of this research. The existing WSO2 machine learner [25] works pretty well. However, it does

not have the ability to give an accurate prediction using streaming data (online data). Our main objective is to develop a machine learner which can give accurate results using streaming data. Here we are mainly going to consider about malware family classification in HTTP data stream.

In order to achieve the goals of the proposed project, there are 4 major tasks to be completed. Those are,

1. Incremental learning component
2. Predictive model with HTTP data streams
3. Data visualization component
4. Implement Ensemble methods.

Creating the Incremental learning component

In our proposed solution we create a feature table which contains the most common parameters in an HTTP request. We capture the HTTP data stream, and essential data to fill the feature table is extracted from the data stream.

Extracted data is cleansed and divided as training and testing dataset to feed the feature table. We input training dataset and train the machine learner to analyze the feature table and to identify most common values for parameters.

The architecture should identify patterns in the history and update the patterns with incoming data without catastrophic forgetting. To accomplish this, the proposed research will create the incremental learning component.

The idea of incremental learning with streaming data focuses on two objectives:

1. Identifying patterns in the recent history.
2. Updating the patterns with incoming data without catastrophic forgetting.

Furthermore, the proposed new Machine Learner will automatically detect odd data and remove it from processing. Therefore it will result in high- accuracy and best quality output.

Predictive model with streaming data

Most common algorithms to classify malware families are Naïve Byes [3-8], SVM [9-16] and Random Forest Algorithm [17-19].

Naïve Byes Algorithm

Naive Bayes [3-8] is an eager learning classifier, and it is sure fast. It used for making predictions in real time.

It has the multi-class prediction feature. Here we can predict the probability of multiple classes of the target variable.

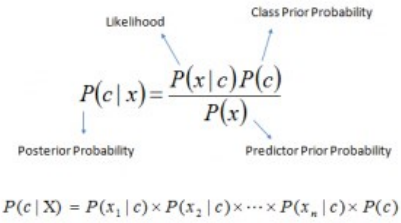


FIGURE II. NAÏVE BAYES ALGORITHM

SVM Algorithm

Support Vector Machine" (SVM) [9-16] is a supervised machine learning algorithm which can be used for both classification and regression challenges. We can plot each data item as a point in n-dimensional space (where n is a number of features you have) with the value of each feature being the value of a particular coordinate. By finding the hyper-plane that differentiates the two classes we perform classification.

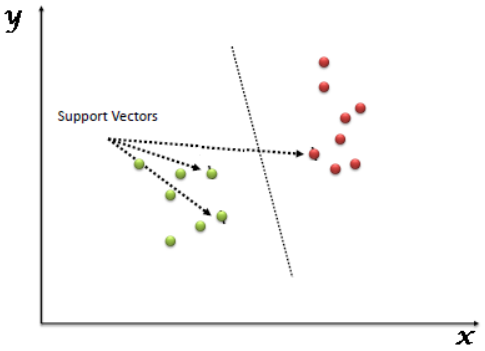


FIGURE IIIII. SVM ALGORITHM

Random Forest Algorithm

Random Forest [17-19] is a versatile machine learning method capable of performing both regression and classification tasks.

We grow multiple trees as opposed to a single tree in CART model in Random Forest. To classify a new object based on attributes, each tree gives a

classification, and we say the tree "votes" for that class. The forest chooses the classification having the most votes (over all the trees in the forest) and in the case of regression, it takes the average of outputs from different trees.

FIGURE IV. RANDOM FOREST ALGORITHM

Current WSO2 machine leaner is capable of classifying data using Naïve Byes [3-8], SVM [9-16] and Random Forest [17-19] Algorithms. Once the machine learner is trained testing dataset is used to classify the malware family. By using testing dataset, we identify whether there are any abnormal behaviors in the feature table. If we identify any abnormalities in feature table via our proposed machine learner, we detect the malware family and predict the malware family classification. Malware family is classified using above three algorithms.

According to the results of this predictive model, the performance of the machine learner will be increased with online streaming data for malware family classification.

Data visualization component

Once the data is processed, it will appear to the user in a format which any user can understand. Data will be presented in a way anyone can easily understand with a basic computer knowledge and basic English knowledge. Proposed system uses multiple visualizations to explore the data such as scatter plots, histograms, Trellis charts, cluster diagrams and so on.

Implement Ensemble methods

The whole idea is to employ multiple apprentice methods and combine their predictions. The aim set of methods is the combination of several predictions based estimators constructed as a learning algorithm in order to improve the generalization / robustness on one estimator. It provides better predictive accuracy than a single learning algorithm. The final prediction will be done by taking the weighted vote of the predictions of the gathered algorithms [20-24].

Naïve Bayes algorithm [3-8], Random Forest algorithm [17-19] and Support Vector Machine (SVM) algorithm [9-16] exist in WSO2 Machine Learner is used to get the Malware family classification of the HTTP data stream.

After data is processed through these algorithms and an outcome is predicted, Ensemble methods[20-24] for Machine Learning is implemented to combine these multiple algorithms such that a better predictive accuracy could be achieved than what could be achieved from a single learning algorithm[10]. Final prediction is done by taking a weighted vote of the predictions of the combined algorithms.

V. FUTURE WORK

By improving the proposed research, we are planning to identify the malware rather than identifying the malware family. It will be a huge task and the team is expecting to focus on some specific malware in order to identify them. Furthermore by improving this research work we can use to bio medical science as well. We can improve this system in order to identify changes in DNA and genes.

VI. CONCLUSIONS

The Hypertext Transfer Protocol (HTTP) has become a universal transport protocol. Timely and accurate detection of anomalies in massive HTTP data streams plays a major role in preventing machine failures, intrusion detection, and dynamic load balancing. Advanced malwares are posing a severe threat to the internet and computer systems. We considered the application of techniques from machine learning, data mining to the problem of detecting and classifying unknown malicious on http data streams. For the malware classification we are use three algorithms such as Naïve Bayes algorithm [3-8], Random Forest algorithm [17-19] and Support Vector Machine (SVM) algorithm [9-16]. Finally Ensemble method [20-24] implemented to obtain accurate prediction. We hope that such a strategy for detecting and classifying malicious will improve the security. We can improve these results in the future.

REFERENCES

[1] [Online]Available: https://www.symantec.com/avcenter/reference/striker.pdf [Accessed: March 8th, 2016]

[2] [Online]Available: http://www.computerhope.com/jargon/p/polyviru.htm [Accessed: March 8th, 2016]

[3] [Online]Available: https://en.wikipedia.org/wiki/Naive_Bayes_classifier [Accessed: March 8th, 2016]

[4] [Online]Available: http://software.ucv.ro/~cmihaescu/ro/teaching/AIR/docs/Lab4-NaiveBayes.pdf [Accessed: March 8th, 2016]

[5] [Online]. Available: Constrained Clustering, Sugato Basu, Ian Davidson, Kiri L. Wagstaff [Accessed: March 8th, 2016]

[6] [Online]Available: http://users.ics.aalto.fi/juha/papers/TRUSTCOM15_Android.pdf [Accessed: March 8th, 2016]

[7] [Online]Available: http://www.aicit.org/JCIT/ppl/JCIT%20VOL7NO5_part25.pdf [Accessed: March 8th, 2016]

[8] [Online]Available: http://www.analyticsvidhya.com/blog/2015/09/naive-bayes-explained/ [Accessed: March 8th, 2016]

[9] [Online].Available: Support Vector Machine, Jason Weston [Accessed: March 8th , 2016]

[10] [Online].Available:Making Large-Scale SVM Learning Practical,Thorsten Joachims[Accessed: March 1st, 2016]

[11] [Online].Available: Support Vector Machines and Metamorphic Malware Detection,Thorsten Joachims[Accessed: March 1st, 2016]

[12] [Online]Available: http://www.svm-tutorial.com [Accessed: March 8th, 2016]

[13] [Online]Available: http://www.iis.sinica.edu.tw/page/jise/2015/201505_11.pdf [Accessed: March 8th, 2016]

[14] [Online]Available: http://phdthesis.uaic.ro/PhDThesis/Cimpoe%C8%99u,%20Mihai,%20%20Classification%20Algorithms%20for%20Malware%20Detection.pdf [Accessed: March 8th, 2016]

[15] [Online]Available: http://www.ijcsit.com/docs/Volume%206/vol6issue04/ijcsit2015060424.pdf [Accessed: March 8th, 2016]

[16] [Online]Available: http://scholarworks.sjsu.edu/cgi/viewcontent.cgi?article=1405&context=etd_projects [Accessed: March 8th, 2016]

[17] [Online]Available: https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm [Accessed: March 8th, 2016]

[18] [Online]Available: https://arxiv.org/ftp/arxiv/papers/1205/1205.3062.pdf[Accessed: March 8th, 2016]

[19] [Online]Available: https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm [Accessed: March 8th, 2016]

[20] [Online].Available:https://www.toptal.com/machine learning/ensemblemethods-machine-learning [Accessed: March 3rd, 2016]

[21] [Online].Available: http://in.mathworks.com/help/stats/ensemblemethods.html?requestedDomain=www.mathworks.com [Accessed: March 1st, 2016]

[22] [Online]. Available: https://datajobs.com/data-science-repo/Ensemble-Methods[Lior-Rokach].pdf [Accessed: March 8th, 2016]

[23] [Online]. Available: Ensemble Learning, Thomas G. Dietterich [Accessed: March 8th, 2016]

[24] [Online].Available: http://web.engr.oregonstate.edu/~tgd/publications/mcsensembles.pdf [Accessed:March 8th, 2016].

[25] [Online].Available: https://docs.wso2.com/display/ML100/Introducing+Machine+Learner [Accessed:March 8th, 2016].

NExT-Plugin for MetaTrader 5 based on Recurrent Neural Networks

H.A.C.J Hettiarachchi, L.I Hettiarachchi, M.M.N.M.Mullegama and R.K.T.D Ranaweera, L.Rupasinghe, D. Kasthurirathna

ABSTRACT

This paper reports evidence that a recurrent neural network model applies to the prediction of foreign exchange rates in the Forex Market. In present human brain power has been used to perform the complex process. It essentially needed an effective and efficient method to replace this exciting procedure in order to improve performance and support time serious analysis. Historical and Streaming Time Series data and technical indicators, such as the impact of economic events, are fed to the neural network to capture the underlying "rules" of the movement in currency exchange rates. The variation of currency pair EUR/USD are forecast by the trained recurrent neural network. The results presented here show that without the use of extensive market data or knowledge, useful prediction can be made, and significant profits can be achieved with reliable predictions. Further research on improving the performance using distributed computing is also conducted. However, the experiments show that with the efficient market it is not easy to make profits using technical indicators or time series input neural networks. After presenting the experimental results, a discussion on future research concludes the paper.

***Keywords*—forex; recurrent neural network; MetaTrader; time series; forecasting;**

1. INTRODUCTION

Foreign exchange market also known as Forex Market is the largest and most volatile financial market in the world and exchange rates are the most important economic indices in Forex Market. Conventional statistical and econometric models used for predicting in the field of financial time series fails to efficiently handle vague nature of foreign exchange data series. Exchange rates are influenced by a number of economic and political factors hence predicting exchange rates is a difficult task from both theoretical and practical point of view. During the time, many statistical and econometric models have been developed by researchers for the purpose of forecasting exchange rates but this problem remains one of the major challenges in the field of forecasting methods. Researches have shown the prediction power of the artificial neural networks. It has been demonstrated that a neural network can approximate any continuous function. Neural networks have the advantage that can approximate any nonlinear functions without any appropriate information about the properties of the data series. In this paper, we present the performance of a recurrent neural network for predicting the exchange rate of EUR/USD. [1] We used data series with daily exchange rates starting from 2011 until 2015 provided by HisData.com.

2. FOREX MARKET

Forex Market is a global network of computers and brokers where traders can buy, sell or exchange currencies. Traders all over the world connect through computer networks, and transactions are executed. As the average turnover of the Forex market is increasing continuously nowadays most of the traders tend to use automated trading, [2] where trading is executed by a software program which is based on a certain trading strategy. Traders make use of various tools for exchange rate prediction in trading. Most of these automated tools are based on artificial neural networks which will predict the future variation of exchange rates.

3. METATRADER 5

MetaTrader 5 (also known as MT5) [3] is an online trading platform developed by a Russian software company MetaQuotes. [4] MetaTrader 5 comes with a built-in editor and a scripting language MQL5 for developing software for financial markets. MetaTrader 5 is the trading platform designed to arrange brokerage services in Forex, CFD, Futures, as well as equity markets. With this platform, you can offer your clients the best trading terminal and better services, thus enabling them to execute trades in several financial markets and stock exchanges in one account. Custom plugins can be installed in MetaTrader 5 and traders can automate the trading process using Expert Advisors. MetaTrader 5

supports for Windows operating system as well as Android and iOS mobile platforms.

After installation, the users can start MT5 app to trade using two options. Beginners can start with demo accounts, and experienced users can obtain registered accounts for trading. Users can add money from the e-money sources. In demo accounts, users can user virtual free demo money for investment but they can't withdraw.

3. METHODOLOGY

Choosing the Neural Network Architecture

As a result obtained from experiments that can be carryout in order to identify a most suitable approach for predictions. Recurrent neural network result is highlighted due to its accuracy, high performance, and capability of handling time series analysis. Recurrent neural networks are used for the scenarios that need the data of the previous state. If there's a need to train a neural network to predict what the exchange rate is would be in the next frame, it would be helpful to know what the rate was in the last frame. Recurrent neural networks seem to be the best choice in forecasting of financial time series. With recurrent neural networks, we take advantage of the modeling abilities of the neural networks for time series forecasting. The data passing through the neurons leaves effective data in the network. Then that memory will be used for the predictions.

Structure of the Recurrent Neural Networks

Neural networks have hidden layers. Usually, the output of the hidden layer is based on the input layer.

Input → hidden → output

However, in recurrent neural networks, the output of the previous hidden layer becomes an input of the previous hidden layer.

Example:

Input-1 → hidden-1 → output

(Input-2 + hidden-1) → hidden-2 → output

(Input-3 + hidden-2) → hidden-3 → output

By looking at the last line, we can see that the 'hidden-3' is affected by every input so far. This is called as the hidden recurrence. Each row of the input data is used to generate the hidden layer (forward propagation) , and then the hidden layer is used to generate the output (Assuming there is only one hidden layer).

Neural Network Memory

Example: (Assuming there are only 3 inputs and 1 hidden layer)

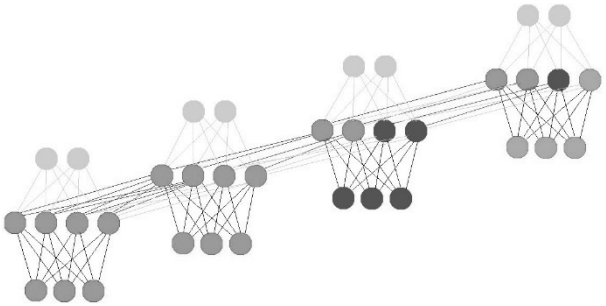


Figure 8-Time steps of recurrent neural network

Above picture reflects 4 time steps. The first time step is only based on the input data. The second time step is based on the input data as well as the hidden layer of the first time step. This continues on. In the 4th time step we can see that the hidden layer memory is full. In the next time step it would have to decide which memories to keep and which memories to overwrite. Bigger layers can hold more memories for a longer period. This is where the neural network learns to remove irrelevant memories and keep important memories.

Backpropagation Through Time

After fully propagating from 1 to 4 we can have the error respect to the value of the output layer and the actual value. And then starts back propagating from 4 to 1.

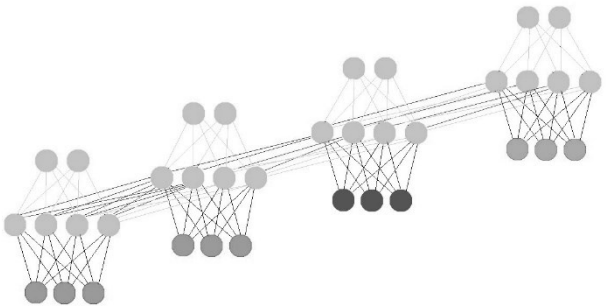


Figure 9-Back propagation

Using Recurrent Neural networks for Forex Market Predictions

The RNN is structured in such a way that it has 9 inputs, 36 hidden neurons with two hidden layers and one output. 9 consecutive moving averages were used as inputs, and the 10th value was used as the output. The activator function used here is the sigmoid.

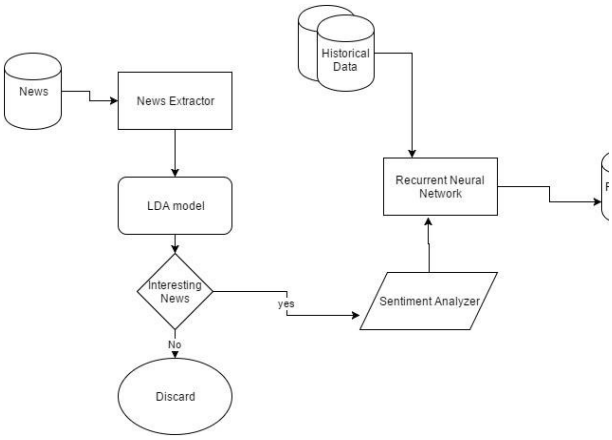


Figure 10-Fine Tuning Prediction

Fine tuning the prediction with the effect of Economic News

News articles are extracted from 4 of the leading financial news sites daily. The news sites are Bloomberg, CNN, DailyFX, FXStreet. Not every News article will affect a currency pair. Since the fluctuations of EURO/USD are predicted, news that has an effect on EURO/USD are filtered out. Therefore, using an LDA model [5] each news article's topic distribution is obtained to filter out most relevant topics to consider.

NExT identifies relevant sentences from news articles by filtering through a keyword list and then uses customized sentiment dictionaries to calculate the value of the sentiment of an article. In this system, the Loughran-McDonald financial dictionary [6] is used to identify relevant keywords about used in financial context. For identifying unanticipated events, the AFINN dictionary [7] is utilized which is commonly used to measure large-scale general emotions.

Sigmoid function

$$S(t) = \frac{1}{1 + e^{-t}}$$

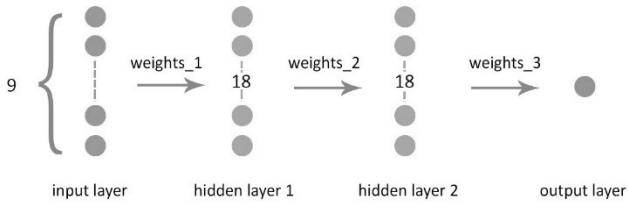


Figure 11-Weights propagation in hidden layers

As shown in Figure 3, weights_1 is a 9x18 matrix which propagates input data to the hidden layer 1. Weight_2 is an 18x18 matrix which propagates the hidden layer 1 data to hidden layer 2. Weight_3 is an 18x1 matrix which propagates the hidden layer 2 data to the output.

Application of Recurrent Neural Network

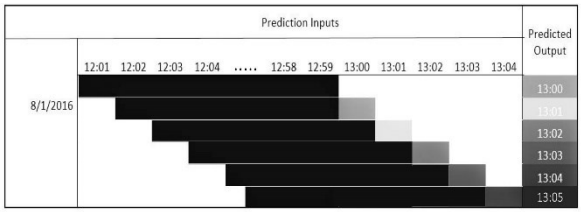


Figure 12-Applying Recurrent Neural Network for Prediction

As shown in Figure 5 above, the algorithm is taking sixty continuous time frames as inputs with the and passing them through the developed neural network to predict the next time frame. The first algorithm considers first the 60 values and predicts the 61st value. Then it starts with the 2nd value and considers the newly predicted value as the last value. After the prediction, data will be saved to the.CSV file for future usage. With the new predicted output from the neural network, it runs repetitively. As considerable time is taken for training the neural network, only fifteen to thirty-time frames from the beginning are trained. Then neural network with sentiments from live news feeds is re-trained to improve the reliability of the prediction.

4. DATASET

Historical data [5] (to be used later on neural network training) has been obtained through registration with a forex broker and as completing our first milestone, all the data from year 2011 to 2016 has been taken into consideration. Dataset contains columns such as date, time, high, open, close and low values respectively. But after going through the dataset, it was identified

that there are some data missing for some time periods. Therefore, Python scripts have been used to check a number of missing records. After identifying a number of missing records, an algorithm was developed to fill the missing values. After doing a good literature review, it was found out that the best way to go is through average values [6].

$$\text{Value}(n_i) = \frac{\text{Value}(n_{i-1}) + \text{Value}(n_{last})}{2}$$

Where,

Value (n_i) → missing value to find,

Value (n_{i-1}) → value before missing value

Value (n_{last}) → first occurrence of value after missing value

By the algorithm, it was found out that the filling values does not affect available data because it always provides an average value.

Table 4-Filling Missing Values

Value	Calculation	Answer
1.6845	-	1.6845
N1	(1.6845+1.7458)/2	1.7152
N2	(1.7152+1.7458)/2	1.7305
N2	(1.7305+1.7458)/2	1.7381
1.7458	-	1.7458

After filling missing values, next main issue was to save streaming data of the Forex market. (Since, as the time passes streaming data become historical data). Therefore it was decided to use MetaTrader 5 itself to save its streaming data on-the-go. By using MQL5 language on MetaTrader 5 its live data feed was accessible to obtain information about data and time, high, low, open value and close value. All historical data was stored in.CSV format (Comma Separated Version) to make it easy to feed in to the neural network.

After successfully saving live data feed, next we economic calendar data was obtained from DailyFX [7] to fine tune our predictions. Exchange rates of currency are affected by economic and political events that are taking place within the country. [8] Therefore, it is important to take the impact of these events into consideration when forecasting exchange rates. Impact level of an event is indicated as high, medium and low. Using information of impact level, it was decided how does the prediction value vary.

5. CONCLUSION

In this paper, it is been discussed how to use recurrent neural networks in Forex prediction. NExT plugin is an attempt to bridge the currently used primitive techniques on Forex trading all over the world. Using NExT Forex traders can get a prediction so that they can invest wisely.

However, a user with a significant level of understanding of Forex market, can improve and develop their ability of trading into more advanced trading. NExT plugin will secure you by predicting the best time to buy or sell.

Currently available solutions provide users predictions based on statistical and machine learning approaches. NExT is based on deep learning techniques which are new to MetaTrader 5. There are few drawbacks on first stage of training of the neural network but after the initial training it can be used for several weeks.

Predictions are further fine-tuned by considering the impact of External Economic factors as well, which is a novel feature of NExT plugin. Machine learning was used to integrate impact of external events on trading increasing reliability. As future development we hope to implement a distributed architecture to increase the performance using parallel processing.

REFERENCES

[1] "World Currencies," [Online]. Available: http://money.cnn.com/data/currencies/.

[2] "Autotrading," [Online]. Available: http://www.investopedia.com/terms/a/autotrading.asp.

[3] "Meta Trader Trading Platform," [Online]. Available: http://www.metaquotes.net/en/metatrader5.

[4] "Home," [Online]. Available: http://www.metaquotes.net/.

[5] "Forex Data," 2012. [Online]. Available: http://www.histdata.com/download-free-forex-data/ .

[6] K. Taghandiky, "Research Gate/What are the best methods for filling in missing values," 23 5 2015. [Online]. Available: https://www.researchgate.net/post/what_are_the_best_methods_for_filling_in_missing_values .

[7] "Economic Calendar," [Online]. Available: https://www.dailyfx.com/calendar.

[8] J. V. Bergen, "6 Factors That Influence Exchange Rates," [Online]. Available: http://www.investopedia.com/articles/basics/04/050704.asp.

[9] "BabyPips," [Online]. Available: http://www.babypips.com/tools/forex-calendar/.

[10] "Economic Calendar," [Online]. Available: https://www.dailyfx.com/calendar.

[11] "Foreign exchange market," [Online]. Available: https://en.wikipedia.org/wiki/Foreign_exchange_market.

[12] "MetaTrader 4 for pc," [Online]. Available: http://www.forextime.com/forex-

trading/platforms/download-metatrader4-mt4-windows.

[13] "Range Bars Plugin," [Online]. Available: http://www.az-invest.eu/rangebars-plug-in-for-metatrader4.

[14] R. Aamodt, "Using Artificial Neural Networks To Forecast Financial Time Series," 2010.

[15] M. Kotyrba, E. Volna, D. Brazina and R. Jarusek, "ELLIOTT WAVES RECOGNITION VIA NEURAL NETWORKS," Ostrava.

[16] A. A. Philip, A. A. Taofiki and A. A. Bidemi, "Artificial Neural Network Model for Forecasting Foreign Exchange Rate," *World of Computer Science and Information Technology Journal (WCSIT),* vol. I, no. 3, pp. 110-118, 2011.

[17] R. H. Shumway and D. S. Stoffer, Time Series Analysis and Its Applications, 3rd Edition ed., New York: Springer, 2011.

[18] T. Vanhove, G. V. Seghbroeck, F. D. Turck, B. Vermeulen and P. Demeester, "Tengu: an Experimentation Platform for Big data Applications".

[19] A. v. Waldburg, "FRACTAL WAVES in STOCK MARKET PRICES," 2007.

[20] B. Oancea And S. C. Ciucu, "Time Series Forecasting Using Neural Networks".

Mıtıgatıon of The Evıl Influence of Internet Use On Teenagers in Srı Lanka

T.G. Premarathne, Yasas Mallawaarachchi

ABSTRACT

This is the information era. Whole the world is considered as a single village alias the concept of the global village has been accepted. Because each and every country is connected via Information and Communication Technologies (ICT).

Specially, it is the Internet which is a connection of millions and millions of computers or servers throughout the world. It is said that more than 100 countries are connected into exchanges of data, information, news, knowledge, opinions and many more over the Internet. So, the Internet plays a main roll on the concept of globalization. Hence, the use of Internet is experienced by everyone everywhere and every time.

The Internet has divided the society into two portions called "Digital Natives and Digital Immigrants". With the rapid development of ICT, there is a rapid trend of use of internet among the Digital natives alias the young generation in Sri Lanka, especially among teenagers.

Digital Natives' this new trend of Internet use has brought out another bad trend of evil. So, this evil influence of internet use on teenagers was surveyed.

The survey has proven that the evil influence of Internet use has affected the teenagers' physical and mental health which causes to raise many social and economic issues.

After the study of the research data, the main issues as well as the causes for the evil influence of internet use, for the Mitigation of the evil influence of Internet use on teenagers, suggested recommendations are as follows:

To enhance the parents' observation on the teenagers' internet related activities.

To improve the awareness on issues arise via the Internet use, among the teachers, parents, and the teenagers.

To enhance of the parents' and teachers' guidance to the teenagers on their internet related activities.

1. INTRODUCTION

The Internet is something that affects the life of everyone. Moreover, the life is not easy without the internet today. Specially, the modern life of the young generation depends on the internet. The modern young generation or Digital Natives is embracing the internet thoroughly. The internet has become as the most required equipment of the young generation today.

Education is the main purpose of the adolescents. Moreover, the internet has become as a main source of knowledge. There are all the learning materials on the internet. So, many learning things can be found without any difficulty by a student in anytime anywhere if a student is equipped with the internet.

So, many of young students use the internet, as it is very useful for the success of their education life. No one can reject it. So, all the students have to use the internet for their educational activities.

One more reason to use the Internet by teenagers is leisure. The internet is a paradise for amusement. Some youngsters do the internet sur- fing as a hobby. Some teenagers use the internet as a leisure time activity. Some girls and boys enjoy the internet in watching movies. All type of movies can be enjoyed freely or by paid. There are different types of websites for movies which can be enjoyed freely. There are many torrent sites also for easy film downloading.

The music is also another one of the entertainment among teenagers. All types of music and musicians of all over the world and their creations can be entertained easily by any adolescent through the

internet. Live Television and Radio channels are also can experience on the internet freely.

Now it is getting rare to go out door games and athletics. The new generation is getting used to playing online games. It is not rare to find out any piece of the game over the internet.

Social media is another well-known and charming matter among the young generation over the internet, today. It is a difficult task to find out a teenager who does not have social media account, at least one account. There are many kinds of social media sites, portals, and website on the internet forest. Most of the teenagers apply social media accounts to communicate with their friends and colleagues over the Internet. The trend of finding new friends of the new generation is social media portals over the internet. Most of the adolescents use social media sites to chat with their companions not only by voice chat but also by video. Moreover, it is very simple to create an email account, or any other social media account freely.

Other than these, new generation uses the internet for news, shopping, advertisement, banking and many more.

The atmosphere for use of Internet is wide. Some adolescents use their parents' devices to access the internet. Some have their own devices such as personal desktops, laptops, tabs and smart phones, etc. On the other hand, Information and Communication Technology (ICT) is being rapidly developed. The internet access has been easier and faster than the past. The new technology, known as 4G (Fourth Generation) or LTE (Long-Term Evolution) is used for wireless communication of high-speed data for mobile phones and data terminals such as smart phones.

Broadband technology in Sri Lanka has been expanded its capacity with the concept of FTTH

(Fibre to the Home) technology, taking the speeds up to as much as 100 Mbps.

According to the statistics mentioned in the Annual Report of Telecommunications Regulatory Commission of Sri Lanka (TRCSL) in 2014, it is about 2.5 million of Internet subscribers.

Many of the young generation use smart phones, desktop computers, laptop computers, tabs and such devices which support internet access easily, and such appliances can be found in any house in Sri Lanka today.

A large number of Internet or cyber cafes can be availed in everywhere at a low cost. In such cafes, the internet can be used without any boundaries. Now, computer centers with the internet access facility are available in many schools in Sri Lanka.

A New project called School net 2 under the Broadband Initiative of TRCSL, High speed internet is provided to main schools in Island Wide.

The government aspiration is also to develop the infrastructure facility to be accessed the internet all over the island, specially, among the children. It is reported that about eight hundred (800) Nanasala centers have been already opened and utilized for the use of the internet.

The population of use of internet is increasing day by day because the rapid development of the Information and Communication Technology. Specially, the new generation is embracing the internet quickly and rapidly. As a result of this, the rate of occurrence of disputes based on the use of the internet is increasing rapidly.

There are some solutions that have been taken by authorities, as a remedy to safeguard the teenagers. Though the government has instructed the Internet Service Providers to constrict some pornographic websites, it is not quite enough to save the children from the internet evil. The amount of restricted website is very few. It is less than hundred. However, there are thousands of vulgar websites and portals that can be accessed easily and freely. Sometimes, in one click on a web link leads to a path more than thousands of links for different types of obscene web pages. On the other hand, it is not easy to restrict such sites. Sometimes, the restricted site can be posed in another name and in another way.

2. PROBLEM STATEMENT

Now, the door has been opened completely to be accessed the internet without much difficulty for the young generation in Sri Lanka. There is no any proper door with a proper lock to enter the internet forest safely. However, it is necessary to have pre-prepared successful safeguard system when an adolescent enters into the internet. Because it is the nature, which the children quickly embrace the evil than the good. The younger generation has been

pushed into the internet jungle to benefit the internet. However, there is no any sufficient protection in the internet jungle to save the children by them- selves. There are different types of brutal, dangerous and fearful human animals in the internet jungle. These brutish human animals are always looking forward to their prey. Moreover, there is all manner of traps and pitfalls which cannot be recognized clearly even by a matured person.

A child specially a girl who never goes the boutique to shop without an adult has been dispatched alone to those above unprotected and complicated internet wood to play or study. However, how he or she can protect himself or herself is not an easy achievement. Many of the candidates who face to this interview fail simply.

It is not much difficult to cheat or mislead an adolescent who enters into the internet. A new comer does not have enough experiences about the evil deeds on the internet. Normally, a teenager is not armed with safeguard devices to fight with the internet enemies such as hackers, internet predators and cyber bullers, etc. it is reported that there are about twenty-five thousand (25000) cyber bullies in Sri Lanka amid the internet.

It was reported that Sri Lanka was the first in Google search to search "sex" via Google in last few years. Some boys had used the internet to revenge or wreak from their girlfriends publishing the girls' naked images and videos on social web sites. Some politicians and racists use the internet to cheat and involve teenagers for their activities.

There are various types of Internet-related issues that can be categorized in some areas as mental issues, physical issues, social and economic issues, other issues.

The current position of different types of issues can be understood with the following mentioned information:

"Sri Lanka is the first in Google search searching "SEX" in past few years.

Incidents Reported to Sri Lanka CERT |CC

Table 2.1: List of The Incidents Reported to Sri Lanka CERT |CC

Incidents Reported to Sri Lanka CERT |CC						
Type of Incident	**Year**	**Year**				

	2014	**2013**	**Year 2012**	**Year 2011**	**Year 2010**	**Year 2009**
Fishing	12	8	8	6	6	10
Abuse/Privacy	8	8	8	2	20	10
Scams	12	18	6	3	10	11
Malware	3	2	2	1	5	12
Defacements	56	16	15	20	8	15
Hate/Threat Mail	10	8	6	3	12	8
Unauthorized Access/Attempted	8	11	1	3	10	3
Fake Accounts Social Media	2250	1200	1100	1425	80	
Intellectual Property Violation	3	3	3	5		
DoS/DDoS	6	1	1	1		
Total	**2368**	**1275**	**1150**	**1469**	**151**	**69**

National Child Protection Authority of Sri Lanka says that information of international internet criminals about twenty-five thousand (25000) have been found.

Other than these, many kinds of pornography, suicides, fake love affairs, revenge, kidnappings, violent information, frauds, etc. were experienced in past few years.

3. RESEARCH OBJECTIVES

3.1 Main Objective

The main objective of this study project is "to minimize the evil influence of internet use on teenagers in Sri Lanka."

3.2 Sub Objectives (Research Purpose)

To identify the issues arise over the internet use.

To identify the causes for issues arise over the internet use.

To make recommendations to minimize the rate of the issues.

4. RESEARCH QUESTIONS

The Research Questions of this study survey are as follows:

What are the issues arise via the internet use on teenagers?

What are the causes for issues arise via the internet use?

How to mitigate the evil influence of the internet use on teenagers?

5. RELATED RESEARCHES

It was difficult to find out a complete research study on "Mitigation of the Evil Influence of Internet use on teenagers in Sri Lanka."

However, Mr. Valdemar W. Setzer has published a review paper of the book by Gregory S. Smith (a Brazilian author) on "How to Protect Your Children on the Internet: A Road Map for Parents and Teachers".

It describes and details the dangers that children and adolescents risk when using the Internet, and challenges parents to verify if they got the proper information about them. Thus it draws the attention on some technical and non-technical perspective precautions, tricks and strategies to protect the children.

Further, Mr. Chinthaka Wijewickrama, President of the Federation of Information Technology Industry of Sri Lanka (FITIS), speaks on the title of "How do we protect children from Internet and Facebook."

To protect the children from the internet and face book, he says that as a precaution, computers should not be placed in an isolated area such as a room, where children can use it at their freewill without being seen or supervised. Moreover, under aged children accessing Facebook is not recommended. It is not at all advisable. Furthermore, use of parental control software is also helpful.

Accordingly, with the help of the above-mentioned literature reviews and the other observations, the issues arise via the internet use can be categorized in four areas as follows:

Mantel Issues

Physical Issues

Social Issues

Other Issues

Internet addiction, playing violent games, Risking chatting etc. can be introduced as Mental Issues. Physical Issues includes poor eye appearance, obesity, postural problems etc. Social Issues category includes Pornography/Sexing, online Prostitution, violent information, cyber fraud, drug trafficking, etc.

Other than these issues, Other Issues are being a victim of cyber space, the collapse of education, reduces of personality development, distancing from the family relationship, etc.

In this survey, I studied the prevailing issues in Sri Lanka and simple devices and basic non-technical strategies to protect the adolescents from the evil influence of the internet. Specially, how teenagers, parents and teachers can safeguard and sustain by themselves.

6. POPULATION & SAMPLE

The total population of the teenagers in this research study is two hundred thirty eight thousand and hundred and seven (238,107). It consists all the teenagers in Colombo District of Western Province in Sri Lanka. Furthermore, there are about thirteen (13) Divisional Secretariats in Colombo District. The sample size of this study survey is 60 teenagers in Colombo district. Individuals were studied at data collecting.

7. HYPOTHESIS OF THE STUDY

The formulated hypotheses in this research study are as follows:

There is an association between the lack of parents' observation on teens' internet related activities and the affected.

There is an association between the lack of awareness on issues arise via the internet use and the affected.

There is an association between the lack of parents' guidance on the internet related activities and the affected.

8. SUMMARY OF THE RESPONSES OF THE RESEARCH QUESTIONNAIRE

The responses of the respondents who were inter-viewed individually have been clearly summarized with percentage of the same in the following

presented table that is the able 8.1: Summary of The Responses of The Research Questionnaire.

Table 8.1: Summary of the Responses of the Research Questionnaire

Summary of the Responses of the Research Questionnaire "Mitigation of the Evil Influence of Internet Use on Teenagers"			
Sample Size	60		
Question Number		Amount	%
1. Age Group	13-16	25	42
	16-19	35	58
2. Schooling	Yes	43	72
	No	17	28
3. Gender	Male	34	57
	Female	26	43
4.Internet Use	Yes	43	72
	No	17	28
5. No. of hours per day on			
Education	Less than 2	11	18
	2 to 4	13	22
	More than 4	4	7
Play Online Games	Less than 2	4	7
	2 to 4	10	17
	More than 4	7	12
Social Media	Less than 2	14	23
	2 to 4	15	25
	More than 4	11	18
Other	Less than 2	9	15
	2 to 4	14	23
	More than 4	8	13
6. Device Owners	Yes	37	62
	No	6	10
7. Device (Yes)			
i. PC		11	18
ii.Lap Top		13	22
iii.Phone/Tab		28	47
8.Parents Observation	Yes	10	17
	No	33	55
9. Guidance By			
i. Parents	Yes	6	15
	No	37	
ii. Teachers	Yes	8	13
	No	35	
iii.Other	Yes	0	0
	No	0	0
10.Awareness on			
i. Mental Issues	Yes	7	22
	No	36	50
ii. Social Issues	Yes	17	28
	No	26	43
iii. Physical Issues	Yes	19	32
	No	24	40
iv. Other Issues	Yes	15	25
	No	28	47
11.Social Media Account Holders	Yes	40	67
	No	3	5
12.Use of true info	Yes	29	48
	No	11	18

13.Unknown cyber friends	Yes	40	67
	No	0	0
14. Cheated by cyber friend	Yes	15	25
	No	25	42
15. Plagiarism	Yes	6	10
	No	22	37
16. Access to adults only content	Yes	17	28
	No	18	30
17. Post of adults only content	Yes	6	10
	No	29	48

9. IDENTIFICATION OF THE AFFECTED TEENAGERS WITH THE EVIL INFLUENCE OF INTERNET USE

The responses for the question number 5, 14, 16 and 17 of the questionnaire are very critical.

Table 7.1: Summary of the Responses of the Question No. 5, 14, 16 & 17

Summary of the Responses of the Question No. 5, 14, 16 & 17		
Sample Size = 60		
Question Number	Amount	Category of Issue
Spending more than 4 hours		Mental Issues or Physical Issues
ii.Play online games	7	
iii.Social media	11	
iv.Other	8	
14. Cheated by cyber friend	15	Other Issues
16. Access to adults only content	19	Social Issues
17. Post of adults only content	6	
Total = 36(At least affected by 1 issue, some have been affected by more than 1 issue.)		

As a base, the teenager is a student. The free time of a student is at least 7 or 8 hours per day, excluding school time from morning to evening, (5.00 am to 3.00 pm). Usually, the free time is from 3.00 pm to 10.00 pm. Moreover, it is the sleeping time from 10.00 pm to 5.00 am

According to the evaluation of the above-mentioned table details, 20 students have spent more than four hours per day. It assures that they are being affected by the mental issues such as internet addiction. However, the students who have spent 4 hours on education can be released from this assume because spending time on education purpose can not be assumed as an addiction.

15 students have been cheated by an unknown cyber friend means they have become victims of cyber space. They are preys of cyber predators' or bullers' etc. So they are affected with the issues of other category.

19 students access adult only contents such as pornographic video, audio, text or images etc. Furthermore, 6 teens have published pornographic contents. So they can be assumed as the affected with social issues.

Therefore, 36 teenagers are affecting with issues of all categories which were described earlier. Here, some teenagers are affecting at least with one issue and some are affecting two or more issues.

As well as, they may use the internet in vain. It means, they destroy their valuable time, they corrupt their education, and they themselves tear down their lives.

Thus, in brief, these 36 teenagers can be concluded as the affected with the evil influence of internet use.

10. CONCLUSIONS

10.1 Conclusion on the Research Hypothesis

So, the conclusions on the Hypothesis can be concluded as follows:

For the first hypothesis, (There is an association between the lack of parents' observation on teenagers' internet related activities and the affected.) It could be proven as it is true. The association

between the lack of parents' observation on teenagers on their internet related activities and the evil influence of internet use is positive.

Hence, the evil influence of internet use can be mitigated by increasing the rate of parents' observation on teenagers' internet related activities. Therefore, I'd like to propose, as to increase the level of the parents' observation on teenagers' internet related activities.

For the second hypothesis (There is an association between the lack of awareness on issues arise via the internet use and the affected.), it can be assured that there is also an association between awareness on issues arises via the internet use and the affected.

Thus, the evil influence of internet use can be mitigated by improving the awareness on issues arises via internet use.

And, on the third hypothesis (There is an association between the lack of parents' guidance to teenagers on their internet use and the affected.), it can be confirmed that the association between the lack of parents' guidance to teenagers on their internet use and the affected is true. Here, the evil influence of internet use can be mitigated by improving the degree of guidance by parents to teenagers on their internet use.

10.2 Conclusion on the Research Questions and the Objectives

The 1st research question:

What are the issues arise via the internet use on teenagers?

The 1st sub objective:

To identify the issues arise over the internet use on teenagers.

The identified issues can be categorized into four:

Mental issues: (Internet addiction, moral corruption, Insomnia, play violent games, continuous chatting etc.)

Social issues: (Pornography/Sexing, violent information, suicides etc.)

Physical issues: (poor eye appearance, obesity, postural problems etc.)

Other issues: (Being a victim of cyber space, collapse of education, reduce of personality development, distancing from family relation- ship, plagiarism etc.)

Identification of the issues arise over the internet use were done. It could be proven that there were issues of all categories such as mental, physical and social and other types of issues with what teenagers were affected.

The second research question:

What are the causes for issues arise via the internet use?

The second sub objective:

To identify the causes for issues arise over the internet use.

The identified causes are as follows:

Lack of the parents' observation on teenagers' internet related activities.

Lack of the awareness on issues arise via the internet use among the teenagers, the parents and the teachers.

Lack of the parents' guidance to teenagers on their internet use.

The 3rd research question:

How to mitigate the evil influence of the internet use on teenagers?

The 3rd sub objective:

To make recommendations to minimize rate of the issues.

The overall goal of the research:

Mitigation of the evil influence of internet use on teenagers.

To achieve the main objective or overall goal "Mitigation of the evil influence of internet use on

teenagers," other than the recommendations and suggestions done in the previous chapters, recommendations and suggestions are done as follows:

11. RECOMMENDATIONS

The recommendations as solutions for issues to achieve the main objective, “to minimize the evil influence of Internet use on teenagers” are as follows:

1. To enhance the parents’ observation on the teenagers’ internet related activities.

2. To improve the awareness on issues arise via the internet use, among the teachers, parents and the teenagers.

3. To enhance of the parents’ and teachers’ guidance to the teenagers on their internet related activities.

12. FUTURE WORK

“To minimize the evil influence of Internet use on teenagers in Sri Lanka,” a Spiritual Guidance may help much, especially, since all Sri Lankan teenagers accept and observe their religion. However, in this study, I was unable to perform such study. So, doing a proper research study on the spiritual study can be mentioned as a future study area.

This research study was not mainly focused on a technical view, as an example: use of internet blocking and filtering software, etc. So it is necessary to do an enough study on the technical perspective that helps to mitigate the evil influence of internet use on teenagers.

13. ACKNOWLEDGEMENT

This work has been supported by Sri Lanka Institute of Information Technology, Malabe, Sri Lanka.

14. REFERENCES

14.1 Websites

[1] Valdemar W. Setzer, "HOW TO PROTECT YOUR CHILDREN AND STUDENTS FROM THE INTERNET," in www.ime.usp.br/~vwsetzer, 2010. [Online]. Available: http://www.ime.usp.br/~vwsetzer/how-to-protect-review.html. Accessed: Jun. 18, 2015.

[2] S. Systems, "Internet addiction," 2016. [Online]. Available: http://www.addictionrecov.org/Addictions/index.aspx?AID=43. Accessed: Aug. 18, 2016.

[3] P. T. Staff, "Internet safety for children & teens – what parents should know," Child Development Institute, 2011. [Online]. Available: http://childdevelopmentinfo.com/family-living/kids-media-safety/children-teens-web-internet-safety. Accessed: Jun. 20, 2016.

[4] R. Tamia and V. my complete profile, "Translate," 2013. [Online]. Available: http://retnotamia.blogspot.com/2013/04/top-10-negative-effects-of-internet-on.html. Accessed: May 15, 2016.

[5] Z. Institute, "Major issues facing teenagers: Teen suicide, school shootings, Cyberbullying, Internet addiction, teen hookups, T.V. Violence & teen violence, by Ofer zur, Ph.D.," 1997. [Online]. Available: http://www.zurinstitute.com/teenviolence.html. Accessed: Aug. 18, 2015.

14.2 Books

[6] U. S. Sekaran and R. Bougie, Research methods for business: A skill-building approach, 5th ed. Chichester, United Kingdom: Wiley, John & Sons, 2009.

[7] Y. Jewkes and M. Yar, Eds., Handbook of Internet Crime. London, United Kingdom: Willan Publishing, 2009.

[8] Asoka S. Karunanada, How To Do Research. Tharanjee Prints, 2000.

[9] Professional Management Accounting Institute in Sri Lanka, Business Mathematics (BMT) – Technician Stage Study.

14.3 Reports

[10] Central Bank of Sri Lanka, Annual Report 2011. pp 162-174.

[11] Telecommunications Regulatory Commission of Sri Lanka, Annual Report 2012

[12] Department of Census and Statistics, Sri Lanka, Census of Population and Housing– 2012,

15. APPENDICES I

Questionnaire

Mitigation of the Evil Influence of Internet Use on Teenagers

Age Group : 13-16 ☐ 16-19 ☐
Schooling: Yes ☐ No ☐
Gender : Male ☐ Female ☐
Do you use the internet?
Yes ☐ No ☐
If yes,
How many hours per day do you spend on internet?
Education
Less than 2 ☐ 2-4 ☐ More than 4 ☐
Play online games
Less than 2 ☐ 2-4 ☐ More than 4 ☐
iii. Social Media
Less than 2 ☐ 2-4 ☐ More than 4 ☐
Other
Less than 2 ☐ 2-4 ☐ More than 4 ☐
Do you use internet with your own device?
Yes ☐ No ☐
If yes,
What is the internet access device you own?
PC ☐ Laptop ☐ Phone/Tab ☐
Do your parents observe your internet related activities?
Yes ☐ No ☐
Have you been advised on issues related to internet use by?
Your parents Yes ☐ No ☐
Teachers Yes ☐ No ☐
Other Yes ☐ No ☐
Are you aware on internet related issues such as
Mental issues: (Internet addiction, violent games, continuous chatting etc.)
Yes ☐ No ☐
Social issues: (Pornography/Sexing, violent information, cyber fraud, etc.)
Yes ☐ No ☐
Physical Issues: (eye appearance, obesity, postural problems etc.)
Yes ☐ No ☐
Other issues: (Being a victim of cyber space, collapse of education, reduce of personality development, distancing from family rela- tionship, plagiarism etc.)
Yes ☐ No ☐
Do you maintain social media accounts such as Face Book, Skype, Twitter, WhatsApp, Viber etc.?
Yes ☐ No ☐
Have you used true information in those accounts?
Yes ☐ No ☐
Do you have unknown online friends?
Yes ☐ No ☐
Have you been cheated by any of your cyber friend in any way.
Yes ☐ No ☐
Have you used educational articles pub- lished in the internet, in your academic assignments without citing the original source of the content?
Yes ☐ No ☐
Do you access the adults' only web contents? (Images, video, audio, text etc.)
Yes ☐ No ☐
Have you posted adults only contents to the internet?
Yes ☐ No ☐

Thank you for your help

A Comparative Analysis on Technologies Available for Localization in Indoor Navigation

J.A.D.C.Anuradha Jayakody[1], Shashika Lokuliyana[2],K.Lakshmi Lashini[3],S.A.Rangala[4]

Information Systems Engineering, Faculty of Computing, Sri Lanka Institute of Information Technology, New Kandy Road, Malabe, Sri Lanka

[1]anuradha.j@sliit.lk, [2]shashika.l@sliit.lk, [3]lashinir@gmail.com, [4]samudrarangala@gmail.com

ABSTRACT

Indoor positioning systems have become very popular in past decades. These systems have been successfully used in many applications such as asset tracking and inventory management systems. This paper provides an overview of the existing indoor positioning solutions and attempts to classify different techniques and its usage for the localization. Two typical localization methods are addressed in this research work and studied estimation schemes of triangulation and fingerprinting, so analyzed in detail since it is used in the indoor environment. Furthermore, examine the set of performance matrices, and apply evaluation methods to survey some existing systems. Therefore performance comparisons include accuracy, precision, complexity, scalability, robustness, and cost. The experiment result discusses the above mention performance matrices.

Keywords **– Indoor navigation, RSSI, RFID, BLE, Fingerprinting**

1. INTRODUCTION

Localization is the technique to determine the position of an object or a person with the location. According to researchers Indoor localization system is a system that attempts to find the accurate position of the object inside a building, mall, hospital, and so forth [1]. Due to the popularity of mobile computing simulates extensive research on the localization of person or assets. In the present era of mobile devices, location information is crucial in a wide range of applications. In order to meet the user's need, the location information of persons or assets is required which can be provided by the indoor localization system. The position helps for the navigator to navigating, tracking, monitoring the known and unknown environment. The localization is still a challenge for the researchers to find the proper methods for identifying location in both indoor and outdoor environment due to following mentioned reasons:

- Indoor environment are more complex [2], [3].
- There is a signal interference and reflection inside building [2], [3].
- It is highly dependent on the environment such as the position of objects, the behavior of the person [4], [5].
- An indoor communication link is unreliable and inaccurate [4], [5].

Many indoor location-based algorithms are proposed and implemented by researchers, but design and deployment of a system for obtaining location and spatial information in an indoor environment is a challenging task for several reasons mentioned below:

- User privacy [8], [9].
- Management overheads [8].
- System scalability [10].
- Harsh [9], [10].

GPS based sensors play a major role in outdoor localization but due to poor signal strength it cannot be deployed inside buildings, so the cost of GPS device is high for indoor environment, and GPS signals are not available in the confined environment or high-rise buildings [8], [9], [10]. According to research findings, it requires line-of-sight transmission, clear sky-view for proper working, However there are many options are available for the design of positioning system in indoor environment such as infrared, ultrasound, radio frequency identification (RFID), sensor networks, audible sound, light, colour of walls, Wi-Fi access points, BLE beacon sensors, sensors (Accelerometer, gyro, compass, etc.), Magnetic Field, Low orbit satellites, Camera technology. There are some hybrid approaches which combine advantages of two or more different techniques [4], [10].

2. PROBLEM IDENTIFICATION

Indoor Localization System (ILS) defined as "A system that continuously and in real-time can determine the position of object or obstacle in the indoor environment" [13]. Most of the ILSs work in two steps.

- "Where the person is" [4].
- "How to reach to the target person" [13].

According to the [4], [14] an ILS consist three types of location information challenges as mentioned below:

- The absolute location information system should specify the exact location of the end user. Hence, the map of the typical area should be available as an ILS, so to provide localized information. Therefore as a system, it has large computation overheads and more memory [13].
- The relative position information measures the motion of the individuals and reachability to the destination. In [13] stated that the fixed positions are already stored in ILS and changes are noticed related to the fixed position information. It is helpful in tracking a particular object [13].
- The proximity location information specifies relatively known and unknown infrastructure [13].

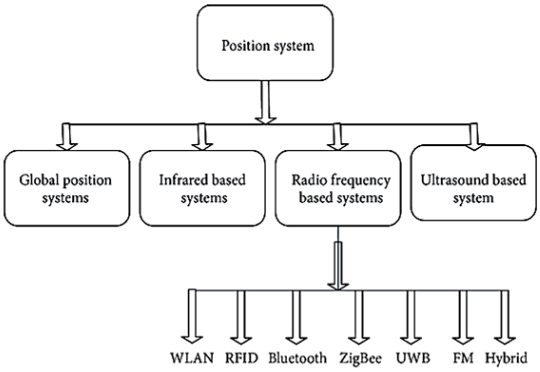


Figure 1: Taxonomy of position systems [14], [16].

Although there are many systems are available as mentioned in Figure 1.This paper addressed both Blue Tooth Low Energy Bluetooth (BLEB) and Wireless systems (WS) as a companied technologies and recommendation based on the performance analysis of the mentioned technologies.

3. RELATED WORK

In spite of the fact that there are advancements, writing survey considered just for the Bluetooth and Wireless frameworks as well-known innovations for the current confinement frameworks.

a. Wireless Local Area Network Based Indoor Localization

J. Padhye, V. Firoiu, and D. Towsley stated that the midrange wireless local area networks (WLAN) are operating in the 2.4-GHz. With a typical gross bit rate of 11, 54, or 108 Mbps and a range of 50–100 m, IEEE 802.11 is currently the dominant local wireless networking standard. Therefore, researchers have been asked to use an existing WLAN infrastructure for indoor localization, by adding a location server [11], [15]. The accuracy of typical WLAN positioning systems using Received Signal Strength (RSS) is approximately 3 to 30 m [11], with an update rate in the range of few seconds. Therefore Wi-Fi positioning systems have become the most widespread approach for indoor localization [15]. Most positioning systems based on WLAN (Wi-Fi) are available as commercial products as prototypes based on measurements on the RSS [11]. Wi-Fi-based positioning systems have two important advantages according to [17], [21], [22].

i. Regarding cost effect, WLAN infrastructures implementation of position algorithms does not need any additional hardware as network interface cards (NICs) measure signal strength values from all wireless access points in the range of the receiver. Therefore, signals needed for positioning can be obtained directly from NICs available on most handheld computing devices. Due to the ubiquity of WLANs, this mode of positioning provides a particularly cost-effective solution for offering LBS in commercial and residential indoor environments [21], [22].

ii. WLAN positioning systems offer scalability in two respects: first, no costly requirement of infrastructure and hardware and second the number of mobile devices subscribing to positioning services [17], [21].

There are several techniques to implement this Wi-Fi positioning system as shown in the Fig. 2 [17]. Authors have studied several triangulation methods under direction and distance based. This paper discussed Signal Property Based Method, under direction based triangulation. The Sri Lanka Institute, the 7th-floor research area, used as a test environment for the experimental results.

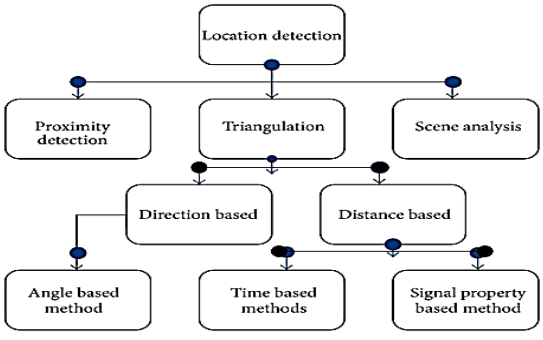


Figure 2: Location detection based classification

1) *Triangulation:* Triangulation uses the geometric properties of triangles to determine the target location. It has two derivations: alteration and angulation. Techniques based on the measurement of the propagation time system (e.g., TOA, RTOF, and TDOA) and RSS-based and received. Signal phase methods are called alteration technique [8], [9]. The AOA (Angle-of-Arrive) estimation technique is also known as an angulation technique.

2) *Signal Property Based Method*: The majority of wireless localization systems compute the distance to the positioning device using either timing information or angle based. In both scenarios, they are influenced by the multipath effect. Due to this, the accuracy of estimated location can be decreased [17]. The substitute method is to estimate the distance of an unknown node to reference node from some sets of measuring units using the attenuation of emitted signal strength [3], [12]. This method can only be possible with radio signals.
Mostly wireless localization systems positioning device using properties of the received signal, with received signal strength indicator (RSSI) being the most widely used signal-related feature. RSSI measurement estimations depend heavily on the environmental interference, and they are also nonlinear [23], [24]. These methods work with the Wi-Fi technology. As this system needs a server for implementation, this technique can work using only access points which are cheaper than Wi-Fi routers.

b. Bluetooth Based Indoor Localization

Bluetooth is a short-range, wireless, cable replacement protocol operates in the 2.4-GHz ISM band [20]. Compared to WLAN, the gross bit rate is lower (1 Mbps), and the range is shorter (typically 10–15 m). On the other hand, Bluetooth is a "lighter" standard, highly ubiquitous (embedded in most phones, personal digital assistants (PDAs), etc.) and supports several other networking services in addition to IP. Bluetooth tags are small size transceivers [20], [25]. Like any other Bluetooth device such as BLEB, beacons has a unique ID. This ID can be used for locating the Bluetooth device. Bluetooth technology is mainly used in small-scale positionings, such as single-room or warehouse. Bluetooth indoor positioning technology of the biggest advantages is the device small and easy to integrate the PDA, PC, and mobile phones, so it is easy to popularize. Indoor localization using Bluetooth beacons is attractive because of the low cost and high spatial selectivity of Bluetooth devices [25]. The beaconing permitted in the BLEB standard enables a very short, unsolicited message at very flexible update rates. These messages can be used to allow a device to detect proximity to a specific location based on the RSS. BLE advertising beacons are particularly attractive to retailers because of the promise of long battery lives of many years, and so low maintenance requirements. In this way, location specific triggers, adverts, vouchers and information can be provided to the user [24], [25]. Long battery lives are expected to require low radio power output and/or low beaconing rates. While this does not affect their use for proximity detection it does affect their usefulness for providing fingerprint-based positioning throughout an entire indoor environment [20] – [25].
Fingerprinting: Fingerprinting is currently the state-of-the-art indoor positioning scheme readily available on standard smartphones [25]. The main theme is to collect features of the scene (fingerprint) from the surrounding signatures at every location in the areas of interest and then build a fingerprint database. The location of an object is then determined by matching online measurement with the closed location against the database [24], [25].

Figure 2: Fingerprinting based positioning [24].

This technique is composed of two phases: Training (Offline) phase and Tracking (Online) phase. During the training phase, signal strengths from Access Points (APS) are collected at pre-identified locations, which are called reference points (RPs). The objective of this operation is building the fingerprint database which will be used in the tracking phase. For the reason that mobile user's location is determined based on the surrounding RPs, they should be distributed in the target area evenly and homogenously [24]. In the tracking phase, Mobile user's surrounding AP RSSs are compared with the RPs dataset collected in the training phase to identify the best matching RPs. The tracking phase could use deterministic and

probabilistic algorithms to match real-time RSS readings with RPs signal data [8], [9], [20] – [24].

4. TESTING AND RESULTS

The examination did on both frameworks which are actualized utilizing Wi-Fi and Bluetooth advances. Formulas:

For Wi-Fi :

$$Distance = 10^{\left[\frac{27.55-(20*\log_{10} frequency)+signal\ level}{20}\right]}$$

For Bluetooth :

$$RSSI\ (dBm) = (-10 * n * \log d) + A$$

d- distance in meters
A-received signal in dBm at 1m
n- propagation constant (n=2 for free space)

In this analysis, the researchers assessed the signal strength of the framework utilizing concentrated on situating strategies. The test outcomes appeared in table 1, and the aftereffect of assessment of execution measurements are portrayed in Table 2.

Table I – Test results

	Signal Strength	
Actual Distance	Wi-Fi (Triangulation)	Beacon (Fingerprinting)
0.1m	-38dBm	-15dBm
1.0m	-41dBm	-26.5dBm
2.0m	-46dBm	-28dBm
3.0m	-49.5dBm	-28.5dBm
4.0m	-52dBm	-28.5dBm
5.0m	-54dBm	-28.5dBm
6.0m	-55dBm	-28.5dBm
8.0m	-58dBm	-32.5dBm
10.0m	-60dBm	32.5dBm

Table II – Comparison of implemented system used for indoor localization

	Systems	
	Wi-Fi	Bluetooth
Accuracy	1 m – 5 m [8]-[10].	2 m–5 m [20]-25]
Principles used for localization	proximity, ToA, TDoA, RSSI Fingerprinting, and RSSI theoretical propagation model [8], [9].	and RSSI theoretical propagation model [20]-[25].
Coverage	Building level [1]-[8]	Indoor [1]-[8]
Power consumption	High [8]-[10]	Low [20]-[25]
cost	Low [8], [9]	High [20], [21]
Remark	1) Infrastructure available everywhere [14], [20]. 2) Initial deployment is expensive [14]-[16].	1) Data transfer speed is high [17],[18]. 2) Limitation in mobility [17], [18].

5. DISCUSSION AND CONCLUSION

This paper reviews the current most prominent indoor situating strategies and frameworks. Diverse execution estimation criteria are talked about, and a few exchange offs among them are watched. For instance, the one amongst multifaceted nature and accuracy needs watchful thought when to pick situating frameworks and procedures for various applications situations, for example, warehousing, mechanical autonomy, or crisis. The assessment of this examination outlines that an area Bluetooth method is better for little ranges while Wi-Fi strategy is reasonable for substantial situations. At long last, this exploration study has demonstrated regarding signal quality, power utilization and information exchange speed Bluetooth situating methods and frameworks altogether expanding the granularity and precision of the area data contrasted with Wi-Fi.

6. ACKNOWLEDGEMENTS

The authors would like to thank Sri Lanka Institute of Information Technology for the support of this research work.

7. REFERENCES

[1] G. Yanying, A. Lo, and I. Niemegeers, "A survey of indoor positioning systems for personal wireless networks," in IEEECommunications Surveys & Tutorials. vol. 11, 2009, pp. 13-32.

[2] G. H. Forman and J. Zahorjan, "The challenges of mobile computing," in Computer. vol. 27, 1994, pp. 38-47.

[3] J. Breckling, Ed., The Analysis of Directional Time Series: Applications to Wind Speed and Direction, ser. Lecture Notes in Statistics. Berlin, Germany: Springer, 1989, vol. 61.

[4] K. Kakousis, N. Paspallis, and G. A. Papadopoulos, "A survey of software adaptation in mobile and ubiquitous computing," inEnterprise Information Systems. vol. 4: Taylor & Francis, 2010, pp. 355-389

[5] G. Gartner, F. Ortag, and H. Huang, "A Survey of Mobile Indoor Navigation Systems," in Cartography in Central and EasternEurope: Springer Berlin Heidelberg, 2010, pp. 305-319

[6] C. Krishna, I. Anand Padmanabha, and N. P. Venkata, "Indoor localization without the pain," in Proceedings of the sixteenth annual international conference on Mobile computing and networking Chicago, Illinois, USA: ACM, 2010, pp. 173-184

[7] N. Bulusu, J. Heidemann, and D. Estrin, "GPS-less low-cost outdoor localization for very small devices," in PersonalCommunications, IEEE. vol. 7, 2000, pp. 28-34. (2002) The IEEE website. [Online]. Available: http://www.ieee.org/

[8] J. Hightower and G. Borriello, "Location Systems for Ubiquitous computing," in Computer. vol. 34, 2001, pp. 57-66.

[9] J. Hightower, B. Brumitt, and G. Borriello, "The location stack: a layered model for location in ubiquitous computing," inProceedings Fourth IEEE Workshop on Mobile Computing Systems and Applications, 2002, pp. 22-28. "PDCA12-70 data sheet," Opto Speed SA, Mezzovico, Switzerland.

[10] T. Camp, J. Boleng, and V. Davies, "A survey of mobility models for ad hoc network research," in Wireless Communications andMobile Computing. vol. 2: John Wiley & Sons, Ltd., 2002, pp. 483-502.

[11] J. Padhye, V. Firoiu, and D. Towsley, "A stochastic model of TCP Reno congestion avoidance and control," Univ. of Massachusetts, Amherst, MA, CMPSCI Tech. Rep. 99-02, 1999.

[12] Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specification, IEEE Std. 802.11, 1997.

[13] J. Hightower and G. Borriello, "Location Systems for ubiquitous computing," in Computer. vol. 34, 2001, pp. 57-66.

[14] Y. Gu, A. Lo, and I. Niemegeers, "A survey of indoor positioning systems for personal wireless networks," IEEE Communications Surveys and Tutorials, vol. 11, no. 1, pp. 13–32, 2009.

[15] D. Zhang, F. Xia, Z. Yang, L. Yao, and W. Zhao, "Localization technologies for indoor human tracking," in Proceedings of the 5th International Conference on Future Information Technology (FutureTech '10), May 2010.

[16] D. Pai, M. Malpani, I. Sasi, N. Aggarwal, and P. S. Mantripragada, "A Robust pedestrian dead reckoning system on smartphones," in Proceedings of the IEEE 11th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom '12), pp. 2000–2007, June 2012.

[17] M. A. Youssef, A. Agrawala, and A. U. Shankar, "WLAN location determination via clustering and probability distributions," in Proceedings of the 1st IEEE International Conference on Pervasive Computing and Communications (PerCom '03), pp. 143–150, March 2003.

[18] C. Pereira, L. Guenda, and B. N. Carvalho, "A smart-phone indoor/outdoor localization system," in Proceedings of the International Conference on Indoor Positioning and Indoor Navigation (IPIN '11), 2011.

[19] J. Wang, R. K. Ghosh, and S. Das, "A survey on sensor localization," in Journal of Control Theory and Applications. vol. 8, 2010, pp. 2-11.

[20] N. Li and B. Becerik-Gerber, "Performance-based evaluation of RFID-based indoor location sensing solutions for the built environment," in Advanced Engineering Informatics. vol. 25, 2011, pp. 535-546.

[21] L. Klingbeil and T. Wark, "A Wireless Sensor Network for Real-Time Indoor Localisation and Motion Monitoring," in International Conference on Information Processing in Sensor Networks, 2008, pp. 39-50.

[22] P. Bahl and V. N. Padmanabhan, "RADAR: an in-building RF-based user location and tracking system," in Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies INFOCOM. vol. 2, 2000, pp. 775-784.

[23] X. Hongli, H. Liusheng, X. Tao, H. Yongkai, and W. Yang, "Practical indoor tracking using Wireless Sensor.

[24] S. N. Razavi and O. Moselhi, "GPS-less indoor construction location sensing," in Automation in Construction. vol. 28, 2012, pp. 128-136.

[25] A. Patil, J. Munson, D. Wood, and A. Cole, "Bluebot: Asset tracking via robotic location crawling," in Computer Communications. vol. 31, 2008, pp. 1067-1077.

ClusterMal: Automated Malware Analysis with clustering, anomaly detection and classification of existing and new behavioral analysis

Prabhath Lakmal Rupasinghe [1], Krishnadeva Kesavan[2], Sripa Vimukthi Bannakkotuwa[3], V.V.YY. Wickramanayake[4], M.P.D.H De Silva[5], J.M.D. Fernando[6], K.K.K.K. Sampath[7]

Department of Information Systems Engineering, Sri Lanka Institute of Information Technology (SLIIT), Malabe, Sri Lanka

[1]lakmalr@gmail.lk, [2]deva130989@gmail.com, [3]sripavimukthib@gmail.com, [4]vijiniyw@gmail.com, [5]dinukhasaranga93@gmail.com, [6]malithdinushan@gmail.com, [7]kalhara.sampath@gmail.com

ABSTRACT

ClusterMal is an automated malware analysis system based on the Cuckoo Sandbox technology. Since malware is the Swiss-army knife of cyber criminals and any other adversary to an organization, in these evolving times detecting and removing malware artifacts is not enough: it is vitally important to understand the behavior, context and motivation and goals of the breach. The Cuckoo Sandbox plays a vital role in analyzing malware but has no clustering feature. As a solution, we propose ClusterMal, which is capable of clustering, anomaly detecting and classifying of existing and new behavioral analysis through machine learning. In the proposed method when a suspicious file is thrown at ClusterMal it throws it into several different environments (VMs with different OSs installed) and a detailed result is outlined. If no similar behavior is observed, a new cluster is created along with the report and a signature for future usage. In practical implication, the module is integrated into Cuckoo for command line interactions with providing high reliability and availability.

Keywords — **Malware Analysis, Clustering, Anomaly Detection, Behavioral Analysis, Sandbox Technology, Machine Learning**

1. INTRODUCTION

In this digital age, open source software is becoming the most significant turning point in the software industry. Even though there are a lot of commercial paid licensed software available, users tend to use open source software as they are free, consist high security, and the source code is available for developing better versions, etc. Meantime, malware threats are becoming more organized and hard to detect. In today's technology-dependent world, enormous amounts of data are stored online and offline every minute and can be immediately subject to attacks. IT professionals do their best to protect the warehouses of information, but as they enhance their efforts, so do the attackers.

A significant threat to data security is malware. It is very important to understand malware, their origins, and the behaviors in order to build the most valuable data protection strategy for companies. For this reason, malware researchers and antivirus engineers had to have a way to analyze malware in depth.

Cuckoo Sandbox technology is open source and is used in a lot of commercial products and services around the world. As we described in this paper main goal of ClusterMal is to further develop this cuckoo sandbox with the capability of clustering the generated reports and making a unique signature for clusters, which can be used as an antivirus signature to detect a lot of different but similar behavioral malware with one signature.

The aim of ClusterMal is to provide a good method to the users who trust in open source software, and for the users who have no ability to buy commercial software.

2. RESEARCH OBJECTIVE

The main objective of this research is to implement a new machine learning module in the cuckoo platform. Clustering, anomaly detection, classification of existing behaviors and new behavior analysis is focused here. This research includes developing a module using machine learning in Cuckoo using Scikit-Learn that should be able to cluster all reports according to similar behaviors. When a new sample is analyzed if no similar behavior is observed, a new cluster should be created along with a virus signature for future references.

The importance of this research is there's the ability to identify how malware act in a different environment by referring the cluster detailed report, also this can be used to produce a signature to the

malware which can be helpful to detect it in the future. Currently, the most of the malware analyzing software are not freely available, but a cuckoo is an open-source software. Hence, with this extended version of Cuckoo (ClusterMal) anyone can use this product freely.

3. METHODOLOGY

The solution overview is mainly based on Cuckoo Malware Analysis which is shown in figure 1. This solution is a module for machine learning in Cuckoo framework using Scikit-Learn that should be able to cluster all reports according to similar behaviors. An algorithm can be developed for clustering purposes using machine learning. Once that a clustering exists and a new sample is analyzed, the new report can be assigned to one of the clusters and compared with similar samples. The module should also be able to perform anomaly detection, hence Alternatively, if no similar behavior is observed, a new cluster should be created. It should be possible to choose among several methods to do this. For example, the distance between the clusters could be measured, or a Support Vector Machines (SVMs) could be trained on existing data using the cluster labels. In machine learning, SVMs are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis [21]. After the functionality based on stored analysis data from Cuckoo Sandbox is implemented, the module will be integrated into Cuckoo for command line interaction.

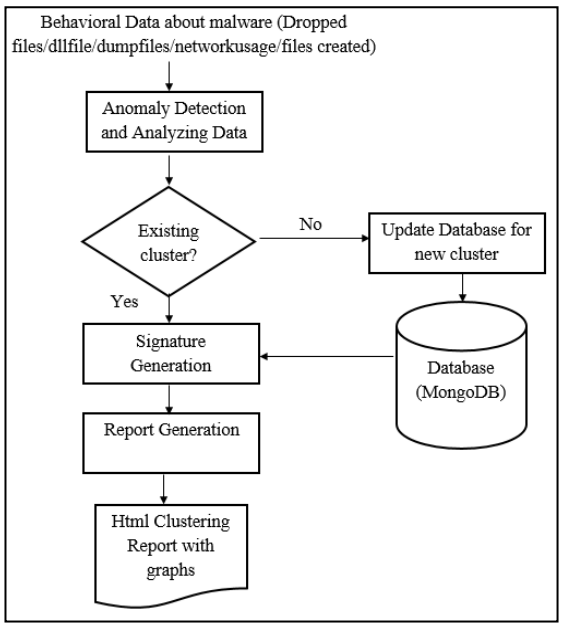


Figure 13: Flow Chart of ClusterMal

ClusterMal is implemented for Linux operating system (Ubuntu 15.0 Wily) based on open-source Cuckoo 2.0 RC1 sandbox. Cuckoo Sandbox is a free open source software that automates the task of analyzing any malicious file under many different environments and produce a report. Python 3.5.1, Yara 3.4.0, g++, MongoDB 3.2, ssdeep and libcap2-bin should be install in the ClusterMal host to run ClusterMal successfully. Few python libraries should be added for several functions as follows,

Python-magic for identifying file formats
Python-dpkt for extracting information from pcaps
Python-mako for rendering html reports and web gui
Python-sqlalchemy
Python jinja2 and Python bottle are necessary for web.py utility

SSDEEP should be installed on ClusterMal host, for calculating fuzzy hashes.libcap2-bin since it is necessary for the cuckoo to run Tcpdump as a non-root user. YARA is used to support identification and classification of malware samples. Python 3.5.1, adobe reader and MS Office should be installed on the VMs. Since ClusterMal has inherited almost all the features of Cuckoo Sandbox, ClusterMal is the new generation of Automated Malware Analysis. Scikit-Learn will be used to develop a module for machine learning in Cuckoo. That would be able to cluster all reports according to similar behaviors.

Research work was carried out to identify available similar applications for malware analysis systems. Linux is used as the operating system and develop using Linux command line. Clustering is used as a method of unsupervised learning, and a common technique for statistical data analysis used in many fields. Here, clustering is the process of grouping a set of objects in such a way that objects in the same group (called a **cluster**) are more similar (in some sense or another) to each other than to those in other groups (clusters).

Scikit is used to learn to develop the malware clustering algorithm. Scikit-learn provides a range of supervised and unsupervised learning algorithms via a consistent interface in Python. It is licensed under a permissive simplified BSD license (a family of permissive free software licenses, imposing minimal restrictions on the redistribution of covered software) and is distributed under many

Linux distributions, encouraging academic and commercial use. There are a lot of benefits in Scikit –learn to develop machine learning algorithms. It provides consistence interface to machine learning models which make is easy to use new models. It also provides many tuning parameters but with the sensible default so that it can be tuned to optimal performance and can use models quickly. It's under active development by active community so it provides the best support. Except for machine learning models Scikit learning also provide a rich set of functionality for companion tasks such as models selection, model evaluation, and data preparation.

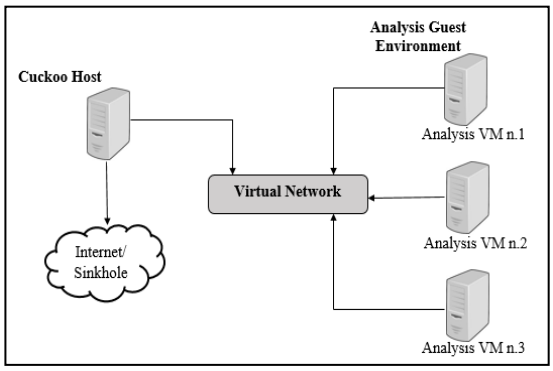


Figure 2: High-level architecture of Cuckoo [15]

Figure 1 and Figure 2 gives the high-level architectures of Cuckoo and ClusterMal where the enhancement of the technology is visible. In this research, a module: ClusterMal is developed for the existing sandbox: Cuckoo and considering the process of extending of the new module in the end to the existing one.

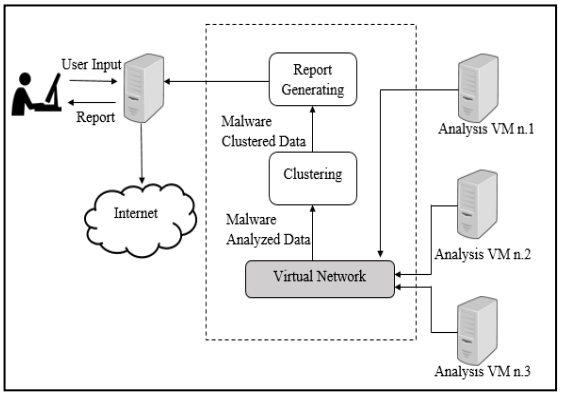


Figure 3: High-level architecture of ClusterMal

The first step is to identify the common behaviors of Malware and Cluster them according to the similar behaviors. After analyzing the data about the behaviors of malware, get the similar behaviors and cluster them accordingly. Then make a cluster of the similar behavior of the malware with the analyzed data. When a cluster exists and a new sample is analyzed, the new report can be assigned to one of the clusters and compared with similar samples. Anomaly detection should be done only if no similar behavior is observed and a new cluster should be created. Also the above process creates a virus signature for each and every clustering groups which is a unique feature and enhances the value of the research. ClusterMal, which is capable of clustering ant it detects anomalies and classify the existing and new behaviors through machine learning.

4. RESULT AND DISCUSSION

Researchers and vendors can get an idea about how the malware acts in a different environment by observing the cluster report. This is an open source approach. Therefore, the solution would help Students who are studying about malware, Antivirus Developers, Malware Analysts and Security Consultants. ClusterMal command line application will only run on Linux Platform. The application will not run on any different operating systems and Linux platforms also should have the recommended software and hardware installed. If anyone needs to install the system first Linux platform is needed.

Under performance requirements, the system will only allow one instance of this application to run on the VM in a given time. Users can not start more than one instance of ClusterMal Application. Multiple users are allowed to use this application from the same VM. The time is taken to generate the report will depends on the size of the malicious file. There will be no specific design constraints to the research. The system will use standard principles of design, and it will maintain the consistency of the user interfaces by using same color and background patterns.

Considering software system attributes high reliability, high availability and high maintainability can be achieved. ClusterMal has less probability of failure since it handle fewer amounts of data but it may fail due to OS failures such as low battery or system crashes. It always provides accurate result because this application contains a machine learning module for clustering. ClusterMal application has high availability. The system installed in user's guest OS will be available at any time.

5. FUTURE WORK

This research can be extended to a level such that it can be more user-friendly. It can be given a new face as an antivirus scanner by using the signatures generated by the ClusterMal. This concept can be further developed as a Security Alert System and send alerts about the currently spreading malware and their behaviors via Email and SMS. ClusterMal can be developed as a web application to enhance the usability. This can be developed for Apple, Android, and Windows platforms as well.

6. CONCLUSION

ClusterMal is an automated malware analysis system using cuckoo sandbox which is capable of clustering, detecting the anomaly and classifying the existing and the new behavioral analysis through machine learning. This is proposed as a solution for the missing of clustering features in the cuckoo sandbox. With the development of this research, there is a facility of identifying similar behaviors of malware, analyzing them and clustering them, when comparing with samples if there are no similar behaviors observed, a new cluster and a signature are created. Using this signature virus guard vendors can implement proper countermeasures for malware in different environments. This research provides a great opportunity to get an idea about how malware acts in a different environment by going through the cluster reports, get the signature details for future detections and have an open source solution for the enhancement in detecting malware.

7. REFERENCES

[1] U. Bayer, I. Habibi, D. Balzarotti, E. Kirda, and C. Kruegel, "A view on current malware behaviors," *Proc. 2nd USENIX Conf. Large-scale Exploit. emergent Threat. botnets, spyware, worms, more*, p. 8, 2009.

[2] U. Bayer, E. Kirda, and C. Kruegel, "Improving the efficiency of dynamic malware analysis," *Proc. 2010 ACM Symp. Appl. Comput. - SAC '10*, p. 1871, 2010.

[3] P. Beaucamps, I. Gnaedig, J. Marion, P. Beaucamps, I. Gnaedig, J. M. Behavior, P. Beaucamps, I. Gnaedig, and J. Marion, "Behavior Abstraction in Malware Analysis - Extended Version To cite this version: Behavior Abstraction in Malware Analysis," 2010.

[4] C. M. Bishop, *Pattern Recognition and Machine Learning*, vol. 4, no. 4. 2006.

[5] F. Channel, S. Area, and A. A. From, "Fibre Channel Storage Area Networks: An Analysis Ins title te Au tho r r eta ins full l rig."

[6] M. Egele, T. Scholte, E. Kirda, and S. Barbara, "A survey on automated dynamic malware analysis techniques and tools," ACM Comput. Surv., vol. 63, no. Ngcit, pp. 41–47, 2011.

[7] S. Gadhiya, "Techniques for Malware Analysis," Int. J. Adv. Res. Comput. Sci. Softw. Eng., vol. 3, no. 4, pp. 972–975, 2013.

[8] E. Gandotra, D. Bansal, and S. Sofat, "Malware Analysis and Classification: A Survey," J. Inf. Secur., vol. 5, no. April, pp. 56–64, 2014.

[9] D. Keragala, "InfoSec Reading Room," SANS Inst., p. 16, 2014.

[10] J. Hegedus, Y. Miche, A. Ilin, and A. Lendasse, "Methodology for behavioral-based malware analysis and detection using random projections and K-Nearest Neighbors classifiers," in Proceedings - 2011 7th International Conference on Computational Intelligence and Security, CIS 2011, 2011, pp. 1016–1023.

[11] H. De Huang, C. S. Lee, H. Y. Kao, Y. L. Tsai, and J. G. Chang, "Malware behavioral analysis system: TWMAN," IEEE SSCI 2011 - Symp. Ser. Comput. Intell. - IA 2011 2011 IEEE Symp. Intell. Agents, pp. 1–8, 2011.

[12] K. Kendall and C. McMillan, "Practical malware analysis," Black Hat Conf. USA, pp. 1–10, 2007.

[13] K. Mathur and S. Hiranwal, "A Survey on Techniques in Detection and Analyzing Malware Executables," Int. J. Adv. Res. Comput. Sci. Softw. Eng., vol. 3, no. 4, pp. 422–428, 2013.

[14] V. Notani, “Behavioral Analysis of Malware,” vol. 787, 2009.

[15] D. Oktavianto and I. Muhardianto, Cuckoo Malware Analysis. 2013.

[16] A. Persistence, “InfoSec Reading Room Sleeping Your Way out of the Sandbox.”

[17] K. Rieck, T. Holz, C. Willems, P. Düssel, and P. Laskov, “Learning and classification of malware behavior,” in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2008, vol. 5137 LNCS, pp. 108–125.

[18] Kromer.pl, ‘Choosing the Best Sandbox for Malware Analysis’, 2013. [Online]. Available: malware-analysis/

[19] Cuckoosandbox.org. [Online]. Available: https://cuckoosandbox.org/index.html

[20] Open-source software [Online].Available: https://en.wikipedia.org/wiki/Open-source_software

[21] Support Vector Machine [Online].Available: https://en.wikipedia.org/wiki/Support_vector_machi

Reduction of Power Losses in Low Voltage Distribution Network by Improving the Distribution Line Connections

Eng. K.Lingaruban, Prof.M.A.R.M.Fernando, Dr. A.Atputharajah, Eng. D.K.P.U.Gunathilake

ABSTRACT

This paper presents attractive technical and nontechnical loss reduction methodologies in an existing Low Voltage (LV) distribution feeder by using refurbishment techniques.

Further, this study describes power losses due to low voltage distribution systems and emphasizes the effects of nonstandard workmanship.

Results were obtained on a selected LV distribution feeder in Jaffna peninsula by using computational tools and reconfigured results were compared with the readings taken before refurbishment.

Keywords – **Loss Reduction, Technical losses, Non-Technical Losses, Electrical Distribution.**

1. INTRODUCTION

An increase in the population and rapid increases in new technology adopted Industries have created high electricity demand in the Jaffna Peninsula, Northern Province, Sri Lanka. After 30 years of long conflict Pease has been established in the northern province of Sri Lanka. The distribution system of Jaffna peninsula has been suffering low reliability and efficiency as the network has not been maintained properly. After the war, a rapid development of infrastructures is now underway, and many investors are opening their activities in Jaffna peninsula. So the demand growth is going to be very high. The existing LV and MV networks are inadequate to meet the present & future demands for Electricity. The distribution networks contribute major quantum of losses in any power system. A considerable amount of energy is lost in the distribution system by the way of technical and non-technical losses. This study is based on Low Voltage distribution feeder in Jaffna peninsula; since the losses are very high in this low voltage distribution network, the technical losses study in selected areas to overcome the current situation. The Laboratory experiments are carried out to determine the losses in the collected distribution line connections from existing low voltage distribution network and the proposed crimping type low voltage distribution line connections.

To check the feasibility and the loss reduction of the proposed work, the annual savings and payback period of the proposed method also determined.

2. DISTRIBUTION SYSTEM LOSSES

The present power systems have requirements to electrify different types of consumer load centers. Such as Domestic loads, Industrial loads, Agricultural loads and general purpose loads. From the generating stations to receiving end several types of losses are exist in the electricity network, Such as Generation losses, Transmission losses, and Distribution losses.

The power losses in electricity distribution network can be divided into two categories. One is Technical losses, and another one is Non-Technical losses or Commercial losses [1],[4]&[8]. Technological and non-technical losses are accounted for 15.52% of total units supplied by selected feeder in an Ilvalai police station area in 2013. Therefore, we have proposed an approach to mainly reduce the technical losses and to investigate the techniques of reducing both types of losses.

2.1 Technical Losses

Every element in a power system (In a line or a Transformer etc.) offers resistance to power flow and thus consumes some energy while performing the duty expected of it. The cumulative energy consumed by all these elements is classified as **''Technical Losses.''** Technical losses are the major part of the electrical losses in the electricity system, resulting in losses in drivers, the iron of the Transformers, eddy currents, connections, corona effect, skin effect and Ohmic losses [1]. Technical losses are related to the material properties it's resistance to the flow of the current through the electrical network causes a higher power loss ($I^2 \times R$) and that is dissipated as heat [1],[2]&[8].

Transformer losses can be divided into two components, namely core loss (no-load losses) and copper loss (load losses). Iron losses are a function of the applied voltage and are often referred to as ‘’no-load losses’’ – they are induced even when there is no load current. Copper losses are functions of the winding current and are often referred to as ‘’load losses’’. These are occurring from the energy required to retain the continuously varying magnetic flux in the core and its invariant with a load on the transformer. Load losses mainly arise from resistance losses in the conducting material of the windings, and it is varies with loading [1].

Line losses:

The major amount of power losses in a power system is in primary and secondary distribution lines.

Therefore the primary and secondary distribution systems must be properly planned and maintained to ensure losses within an acceptable limit. The factors contributing to the increase in the line losses in the distribution systems [4]:

Losses due to overloading:

If we consider a particular distribution feeder the current at tail end of the feeder is low, and the current density in the feeder is increasing towards tail end to the supply end.

To overcome these problems

- To make a limited number of service connections in that feeder.
- Construct a parallel circuit arrangement to distribute the power.
- Using low resistivity material conductors in distribution systems.
- To avoid the use of under size conductors such as to select proper size Conductors.

I^2 R loss

Mainly in the distribution systems huge amount of energy is dissipated by heat (I^2 R Losses).

Methods to reduce I^2R Losses:

Using conductor materials with low resistivity. Using conductors with a large cross section area, such as to use strained multi conductors than using a single steal wire conductor. To reduce distribution line length.

In present CEB system the line length of a distribution feeder is1.8Km. In a feature this line length can be implemented as 1.4Km.

Bad workmanship (Losses in joints and connections): Bad workmanship is resulting in poor contacts at joints and connections in power distribution lines and distribution substations. This contributes significantly towards an increase in distribution losses [10].

At the same time in this context, the following points should be born in our mind. Improper installation and Looseness, Inadequate or poor contact surface area, Using or continuing conventional methods(Mechanical joints and bindings), Loss in over loaded phase wire & neural current flow (unbalance loading), Loosing of Strands, Losses in Service Connection, Undersized service cables, By using conventional methods (Create loose connection in the poles), Use of inappropriate fasteners (without spring washer tightening the bolt & nut in piercing connecter), Losses due to high impedance faults, Unclear way leaves (touching of trees),

Bird nesting in poles, Insulator breaking.

Figure 1: Bad workman practices at the HRFC termination in the selected network

Figure 2: Bad workman practices at the Jumper joint and service connection termination in the selected network

Figure 3: Unclear way leaves at the selected network

2.2 Non-Technical Losses or Commercial losses:

Losses occur on account of non-performing and underperforming meters, wrong applications of multiplying factors, defects in CT & PT circuitry, meters not read, pilferage by manipulating or bypassing of meters, theft by direct tapping, etc. These are all due to none metering of actual consumption and are called **''commercial losses.''**

Commercial losses are caused by non-technical or commercial factors namely pilferage, theft, defective meters, errors in meter readings, estimating the un-metered supply of energetic [2].

Tampering of meters:

Fixing of meters without tamper proof meter enclosure**,** Lack of seal monitoring system**,** Devices are not properly securitized (CTs/PTs are expressed) **Pilferage of energy by illegal tapings:**

Making hooks in overhead bare conductor lines**,** bypassing the neutral wires in meters. **Losses at Consumer End Meters:**

Voltage drop in PT terminals over burned CT

Receipt of Payment:

Lack of system for timely disconnection**,** Lack of knowledge of revenue collection.

Human Error:

Errors in bills and meter readings**,** Lack of a system to deliver bills**,** Avoid meter readings at closed premises.

3. LOSSES REDUCTION TECHNIQUES

3.1 For technical loss reduction:

Re-conductoring of MV lines**.** Construction of new MV backbone lines & gantries, interconnections[8], radial substation, etc, Development of primary substations, Installation of voltage regulators in MV feeders, Feeder reconfiguration**,** Avoid overloading of lines and Transformers**,** Disconnect unloaded transformers**,** Balance the transformers loading**,** Perform regular preventive maintenance**,** Implementing hotline maintenance**,** Static VAR compensation**,** Capacitor bank placement (reduce active power losses)**,** Install booster transformers to long feeders.

3.2 For non-technical losses can easily reduce by: Performing regulator inspection**,** Implementing seal management system**,** Installing tamper proof meter boxes**,** Conducting consumer awareness programmes**,** Implementing substation metering system**,** Installation of static meters (PPM) with high accuracy for Bulk supplies.**,** Remote metering of Bulk supplies.**,** Annual testing of all Bulk supply and 3p60A/30A meters to ensure meter accuracy.

4. LABORATORY TESTING

Laboratory experiments by using the selected samples from the existing network and proposed crimp type of connections.

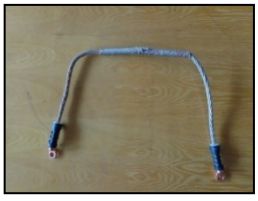
Objectives:-

To calculate the voltage drop and power loss in the collected sample joints from the existing system.

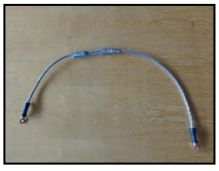
To calculate the voltage drop and power loss in the proposed crimp type joints.

Compression between both.

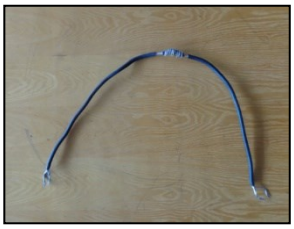
The digital photographs of the collected existing conventional type binding joint samples from the below



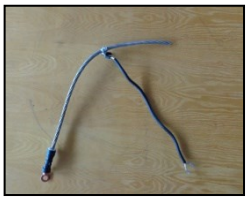
LV Tension joint



LV non tension joint



Service wire mid joint



Service wire termination joint

Figure 4: Collected conventional type electrical binding joints from the existing LV distribution system

The digital photographs of the proposed crimping type joint samples to the selected LV distribution feeder rehabilitation work are given below.

LV tension joint

LV non tension joint

Figure 5: Proposed Crimping type electrical joints

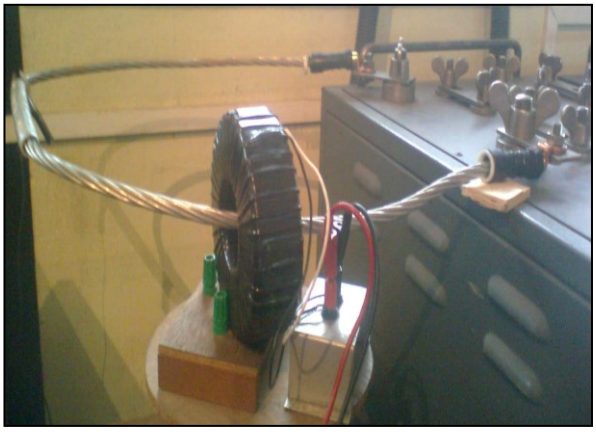


Figure 6: Laboratory testing arrangement for Current Injection, by using current injector

The voltage and current values are measured by using the cathode ray oscilloscope (CRO). The temperature of the joint and the conductor was measured by using Infrared Thermometer. Because when the high current passing through the joint or a conductor, we can't take the temperature readings with the direct teaching method. But for the infrared thermometer temperature measurement method, no need to touch the joint or object physically.

4.1LV Tension joint

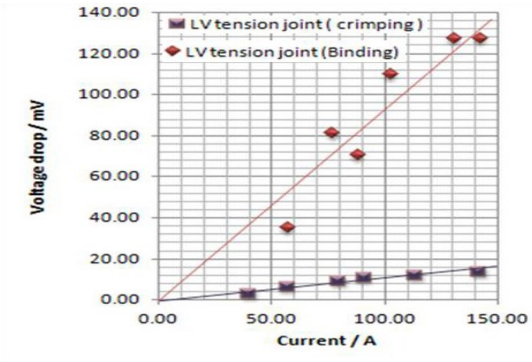


Figure 7: Current Vs. Voltage graph of the LV Tension joint

The current versus voltage drop of the LV tension, joint is given in Fig 7. From this graph it is identified that for the binding joint the voltage drop is increasing with the increase of the current.

However, for the crimping joint, the voltage drop is increasing slightly with the increase of the current. From the comparison between these two graphs, it is concluded that the voltage drop in the LV tension binding joint is much higher the LV tension crimping joint.

From these two graphs, the impedances of the two types of joints were calculated.

|Impedance|= Gradient of the graph

From that, the resistances of the two types of joints were calculated.

Resistance = $|\text{Impedance}| \times \cos(\Phi)$

From that the power loss of the two types of joints also calculated**.**

The power loss in the joint is $I^2 \times R$

Similarly, the power loss in the other type of joints also calculated.

Crimping:

|Impedance|= 0.11 mΩ , Φ=43.2°, Resistance = 0.08 mΩ

Binding: |Impedance|= 0.93 mΩ , Φ=0°, Resistance = 0.93 mΩ

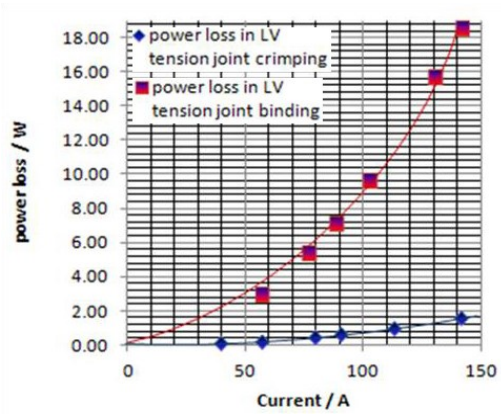


Figure 8: Current vs. Power loss graph of the LV Tension joint

The current versus power loss variation of the LV tension joint is given in Fig 8. From that graph, it is identified that for the binding joint power loss is rapidly increasing with the increase of the current. However, for the crimping joint power loss is varied slightly with the increase of the current. From the comparison between these two graphs, it is concluded that the power loss in the LV tension binding joint is much higher the LV tension crimping joint.

4.2 LV Non-Tension joint

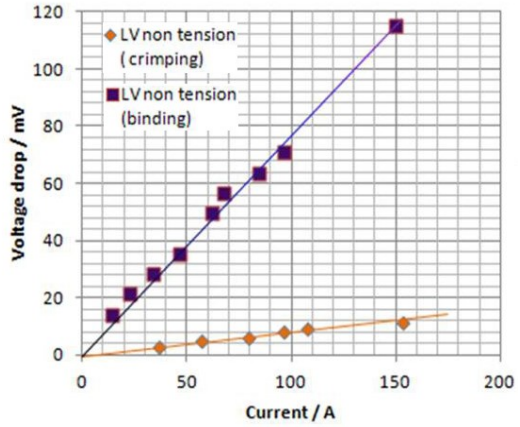


Figure 9: Current vs. Voltage graph of the LV non tension joint

The current versus voltage drop of the LV non tension joint is given in the Fig 9. From this graph it is identified that, for the binding joint the voltage drop is increasing with the increase of the current. However, for the crimping joint, the voltage drop is increasing slightly with the increase of the current. From the comparison between these two graphs, it is concluded that the voltage drop in LV non-tension binding joint is much higher the LV non-tension crimping joint.

Crimping:

|Impedance|= 0.09 mΩ , Φ=36°, Resistance = 0.073 mΩ

Binding:

|Impedance|= 0.77 mΩ , Φ=14.4°, Resistance =0.746 mΩ

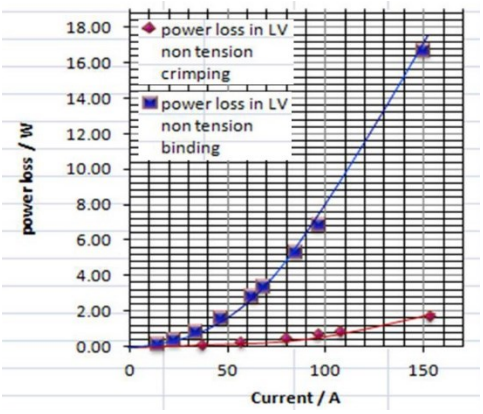


Figure 10: Current vs. Power loss graph of non-tension joint

The current versus power loss variation of the LV non-tension joint is given in the Fig 10. From that graph, it is identified that for the binding joint power loss is rapidly increasing with the increase of the current.

However, for the crimping joint power loss is varied slightly with the increase of the current. From the comparison between these two graphs, it is concluded that the power loss in LV non-tension binding joint is much higher the LV non-tension crimping joint.

4.3 Service Connection Joint

Table 2: Current, Voltage and Temperature readings of the service connection joint both binding and crimping types

Current / A	Voltage drop rms / mV		Temperature °c	
	Binding	Crimping	Binding	Crimping
10	3.56	1.77	29.8	29.8
15	5.67	5.66	29.2	30.7

20	7.07	7.07	30.1	30.7
25	8.48	8.05	30.4	30.7

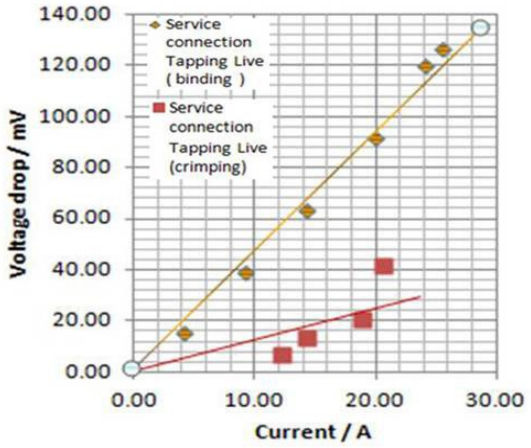


Figure 11: Current vs. Voltage graph of the service connection joint

The current versus voltage drop of the LV service connection joint is given in the Fig11. From this graph it is identified that, for the binding joint the voltage drop is increasing with the increase of the current.

But for the crimping joint the voltage drop is increasing slightly with the increase of the current. From the comparison between these two graphs, it is concluded that the voltage drop in service connection binding joint is much higher the LV service connection crimping joint.

Crimping:

|Impedance|= 1.2 mΩ , Φ=0°, Resistance = 1.2 mΩ

Binding:

|Impedance|= 4.67 mΩ , Φ=0°, Resistance = 4.67 mΩ

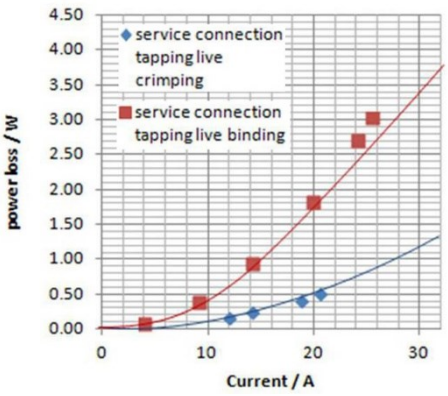


Figure 12: Current vs. Power loss graph of the service connection joint

The current versus power loss variation of the LV service connection joint is given in the Fig 12. From that graph it is identified that, for the binding joint power loss is rapidly increasing with the increase of the current. But for the crimping joint power loss is varied slightly with the increase of the current. From the comparison between these two graphs, it is concluded that the power loss in LV service connection binding joint is much higher the LV service connection crimping joint.

The Summary of the Power loss across the joints at rated current is given below.

Table 3: Summary of the Power loss across the joints at rated current

Sample	**Binding type power Loss / W**	**Crimping type power Loss / W**	**Percentage loss reduction / (%)**
LV Tension joint	9. 6	1.1	88.5
LV Non tension joint	12.6	0.9	92.9
Service Live wire tapping joint	3.4	1.2	64.7

5. CALCULATION AND ECONOMICAL ANALYSIS

According to the energy meter reading from 14/07/2013 to 14/08/2013 the power delivered to the feeder in this period is P_T => 10,397kwh /

month (The energy meter reading at 14/07/2013 - The energy meter reading at 14/08/2013). According to the electricity bill of the consumers in that feeder on the same time period (Please refer appendix A), the power received for the revenue collection is

$P_R = \sum_{i=1}^{100} Pi$ => 8,783 kWh / month.

P_i - Power consumption of a consumer per month.

So the monthly energy loss in that particular feeder is $P_L = (P_T - P_R)$ => 1,614 kWh / month => 2,241.67 W.

6. CALCULATION OF LOSS IN DISTRIBUTION NETWORK BEFORE REHABILITATION

From the laboratory experiments, the total power loss in binding (existing) joints P_{TB} is 1489.2 W. From the laboratory experiments the total power loss in crimping (proposed) joints P_{TC} is 315.3 W.

If we rehabilitate the feeder by using crimping joints the power loss reduction

$P_{LR} = P_{TB} - P_{TC}$ = 1173.9 W

The actual total system monthly energy loss P_L is 1614kWh/month => 2241.67 W

From the laboratory experiments, the total power loss in existing binding joints P_{TB} is 1489.2W. So Non-technical losses + ($I^2 \times R$) losses in the AAC conductor is ($P_L - P_{TB}$) = (2241.67 – 1489.2) =>752.47 W

If we convert all AAC lines by ABC and implement a proper monitoring system for meter sealing, way leaves cleaning & the regular line inspection, we could reduce the non-technical and ($I^2 \times R$)losses $P_{LR50\%}$ minimum by 50%. So loss reduction from non-technical & ($I^2 \times R$) losses in AAC $P_{LR50\%}$ is 376.23W.

Loss reduction from technical losses P_{LR} is 1173.9 W. Total loss reduction per Month is P_{TLR} is=>1550.13 W => 1116.09 kWh /month.

So from this outcome, it is decided the thermal generation of the Jaffna district could be reduced by 1116.09 kWh/month.

During the period 14th, July to 14th August Jaffna district distribution system is totally isolated from the national grid and fed by two thermal generation plants such as Uthuru Janani power plant and Northern power private limited. According to the thermal generation statistics in Statistical Digest 2014, the generation unit price of the Uthuru Janani power plant and Northern power private limited is respectively Rs 18.84 and Rs 34.56. However, according to peak demand data of Uthuru Janani power plant and Northern power private limited, the peak demand is averagely shared by 74.12% from Uthuru Janani power plant and the rest 25.87% from Northern power private limited.

So the generation revenue loss reduction per month is Rs 25563.99.

So the annual generation Revenue loss reduction is Rs 306767.88.

7. ECONOMICAL ANALYSIS

Estimated cost [13] of the feeder Rehabilitation is (See Appendix F & G) =Rs1460000.00 (Additional investment required for the implementation)

Cost saving per year = Rs 306767.88

Time taken to recover the additional investment From the Cost saving is 4.76 Years, Say 5 years.

The life time of the pole line is 25 to 30 years, but from the economic analysis, the payback period is 5 years. Hence this type of loss reduction methodology could be economically acceptable.

8. CALCULATION OF POWER LOSS IN DISTRIBUTION NETWORK AFTER THE REHABILITATION

According to the energy meter reading from 03/06/2015 to 03/07/2015 the monthly energy delivered to the feeder = 11287 kWh / month.

According to the electricity bill of the consumers in that feeder for the time period of 03/06/2015 to 03/07/2015.

The monthly energy received for the revenue collection is,

$$P_R = \sum_{i=1}^{100} Pi \Rightarrow 11051 \text{ kWh / month.}$$

So the monthly energy loss in that particular feeder is 236 kWh / month. But before rehabilitating the feeder the monthly energy loss is 752.47 W =>541.78 kWh / month.

During the cost analysis, we calculated the actual technical loss reduction in Electrical joints and we assumed the balance non-technical losses and ($I^2 \times$ R) losses in AAC could be reduced by 50%. Such as according our assumption the total monthly energy loss after the rehabilitation is (541.78/2) = 270.89 kWh/month.

But after the rehabilitation the actual monthly energy loss in the feeder is 236 kWh/month (<270.89 kWh/month).

So this type of loss reduction methodology can be economically acceptable.

9. CONCLUSION

Power losses in the existing Low Voltage (LV) distribution system network were studied in detail. The power loss in the LV distribution system was studied using data collected from field measurements, laboratory tests and computations. Here the technical and non-technical losses were studied and reduction techniques were analysed in detail. Power losses were checked in several types of joints at the laboratory setups with different materials for different current flow. Critical analyses were done on losses with temperature and current.

The live field measurements of voltage, current and Temperature was taken from the selected LV network with time.

The current versus power loss variation of the LV tension joint is given in the Fig 10. From that graph it is identified that, for the binding joint power loss is rapidly increased with the increase of the current. But for the crimping joint power loss is varied slightly with the increase of the current. From the comparison between these two graphs, it is concluded that the power loss in LV tension binding joint is much higher the LV tension crimping joint. From the comparison between these two graphs, it is concluded that the power loss in LV tension binding joint is much higher the LV tension crimping joint.

Like that it is identified the power loss is increased with current and temperature for the other type of joints.

So the proper type of the electrical joint installation is very much important in distribution system network and this can lead to considerable quantum of energy savings in the distribution system network.

In nutshell, the proposed crimping type joint from this study has been validated as efficient and economical method for Sri Lankan distribution network. This study has proven 65% of loss reduction in laboratory scale as well as in site implementation.

The study economic viability has resulted that implementing the proposed technique on an existing LV network will be 5 years of payback period for a life time of the pole line is about 20 to 25 years.

Hence this type of loss reduction methodology can be economically acceptable.

10. ACKNOWLEDGEMENTS

I wish express my deepest gratitude to, the principal supervisor of this research Prof. M.A.R.M. Fernando, Professor, Department of Electrical and Electronic Engineering (DEEE), University of Peradeniya (UOP) for offering me a chance to work in the area of Loss Reduction studies in Jaffna power distribution system and also his guidance, involvement advices and continuous motivational suggestions including the laboratory testings, without which this research work would not be completed within my targeting time period.

The second supervisor Dr. A. Atputharajah, Dean of the Faculty of Engineering University of Jaffnafor giving his expert knowledge throughout the research works and specially I express my gratitude to Dr. A. Atputharajah for offering me a chance to work in the Department laboratory during weekends. The Valuable hints and ideas gathered from helped me to succeed in this research work.

My Industrial Supervisor, Eng. D.K.P.U. Gunathilake, Deputy General Manager, Ceylon Electricity Board, Northern Province for providing valuable ideas in various situations and supervising my research work towards accomplishment.

I would express my thanks to Engineers, Electrical Superintendents and others who helped me in many ways during the research work period.

I would express my thanks to all the staff members and technical & non-academic staff members of the DEEE, UOP, for helping me in the field and laboratory testing and implementation phase of my research work.

I would especially express my thanks to reviewers and others who helped me in many ways to present this research paper.

Eventually, I may be pardoned if I have inadvertently omitted the names of others to whom my thanks are due.

11. REFERENCES

[1]. A.G.D.P. Jayasekara, J. R. Lucas and W.D. A. S. Wijayapala , Directions for Solving Bad Workman Practices in Power Distribution Line Connections,ENGINEER - Vol. XXXXV, No. 03, pp. [33 - 44], 2012 © The Institution of Engineers, Sri Lanka

[2]. Marwan M. Mahmoud and Imad H. Ibrik ,Power Losses Reduction in Low Voltage Distribution Networks by Improving the Power Factor in Residential Sector. EnergyResearch Centre - An Najah National University Nablus P.O.Box 721 West Bank – Palestine

[3]. N. Rugthaicharoencheep and S. Sirisumrannukul, Feeder Reconfiguration for Loss Reduction in Three Phase Distribution System Under Unbalanced Loading Conditions,Department of Electrical Engineering, Faculty of Engineering, Rajamangala University of Technology Phra Nakhon, Bangkok 10800, Thailand, Department of Electrical Engineering, Faculty of Engineering,, King Mongkut's University of Technology North Bangkok, Bangkok 10800, Thailand, Email: nattachote@ieee.org and spss@kmutnb.ac.th

[4]. L.Ramesh,S.P.Chowdhury,S.Chowdhury, A.A.Natarajan, C.T.Gaunt , Minimization of Power Loss in Distribution Networks by Different Techniques, International Journal of Electrical and Electronics Engineering 3:9 2009

[5]. Y.AL-Mahroqi, I.A.Metwally, A.AL-Hinai, and A.AL-Badi, Loss Reduction and Efficiency Improvement: A Critical Appraisal of Power Distribution Sector in India.World academy of science, Engineering and Technology 63 2012, Reduction of Power Losses in Distribution Systems, Vol.2,Issue.5,Sep-Oct.2012pp-3292-3297, www.ijmer.com

[6]. Solomon Nunoo, Joseph C. Attachie, Frankiln N. Duah, An Investigation in to the Causes and Effects of Voltage Drops on an 11kV Feeder, Department of Electrical and Electronic Engineering, University of Mines and Technology Tarkwa, Ghana,Canadian journal on Electrical and Electronics Engineering Vol.3, No1, January 2012

[7]. Dickson K. Chembe, Reduction of Power Losses Using Phase Load Balancing Method in Power Networks, Member IAENG Proceeding of the world congress on Engineering and Computer Science 2009 Vol I WCECS 2009, October 20-22, 20098,SanFrancisco,USA,www.iaeng/publication/WCES2009-pp492-497.pdf

[8]. Loss Reduction Targets for Region 01 02nd April 2013, NO: DGM (P&D-R1)/DP/Losses/35

[9]. M.Y.Huang, C.S.Chen, C.H.Lin, M.S.Kang, H.J.Chuang and C.W.Haung, "Three phase balancing of distribution feeders using immune algorithm", IET Generation, Transmission and Distribution 2007

[10]. Factors Affecting the Life Time of the Electric Joints M. A. Farahat Electrical Power & Mach. Dept., Zagazig University, Zagazig – Egypt, Proceedings of the 14th International Middle East Power Systems Conference (MEPCON'10), Cairo University, Egypt, December 19-21, 2010, Paper ID 145.

[11]. Catalogue & Price list of materials 2013, Ceylon Electricity Board.

[12]. Tariff system in Ceylon Electricity Board 2013

Engineer adoption of new technologies to boost Social Development; A Perspective on Digitalization of Terrestrial TV Broadcasting in Sri Lanka

Kumudu wasantha Muthuthanthrige, B Sc (Eng), M Eng, C Eng, MIET

ABSTRACT

Seeking continuous improvement in life style possibly would have been the only objective that human race never change across the globe, while others do evolved along time and geographical dimensions.

Discovery of things available in the environment such as fire was the first phase of life style improvement cycle soon replaced by seems never ending second phase of new inventions that takes place even now as we read this paper and will continue until end of civilization.

Clusters of small habitats started networking creating nations along the time axis in the social evolution. Nations having members blessed with genius inventive capabilities started not only enjoying best life styles on earth but also dominate the world maintaining edge over other nations in every respect such as authority, wealth, health and more subtly even how those who are less capable even live.

The inventors became technologically advanced and developed countries those drives improvements in lifestyle others became either followers with some innovations by themselves or lowest category of pure followers setting the so cold development path; begin with adoption then get in to innovation and finally get into elite club of inventers.

The inventors or market drivers collectively make every attempt to discourage others joining their club ensuring always as large as possible market that patronage their new products while individually competing each other by way of introducing different standards to secure a larger individual share of the market. The followers are endlessly facing blinding glare of campaigns of introduction of new technologies as well as subtly differentiating features of competing standards aimed at securing a larger share of the market leaving followers practically lost in a labyrinth of choices in deciding what choice to make and what to avoid; Sri Lanka is known attempting making a choice in selecting a technology standard for Digital Video Broadcasting technology since year 2009 but still seems not yet fully there.

Sri Lanka is the pioneer in South Asia & is a success story in deregulation & adopting new technologies in the telecommunications domain as a result the country enjoys one of the lowest rates & tariffs regime in the world making the services affordable to every cross section of the society, which means Sri Lanka already know benefits of adopting new technologies for a better life.

This paper presents detailed analysis of social benefits of Digitalization of Terrestrial Video Broadcasting Services followed by comparison of technology formats that are available for selection.

The paper discusses an engineering approach in selecting a technology format that considers all relevant aspects such as technological, market share & popularity, end user terminal or TV unit, financial options etc and importance of making available the approach adopted and the final conclusion to the entire society so to make it an inclusive decision than an exclusive decision that naturally leads to disputes and arguments which results delaying adoption and implementation denying all benefits to the country and society. After all economic life of modern technologies is becoming shorter and shorter that demands quick decision making if the society were to reap full benefit package.

1. SOCIAL DEVELOPMENT THROUGH NEW TECHNOLOGIES

Seeking continuous improvement in life style possibly would have been the only objective that human race never change across the globe, while others do evolve along time and geographical dimensions the telecommunications domain. As a result, the country enjoys one of the lowest rates &

tariffs regime in the world making the services affordable to every cross-section of the society, which means Sri Lanka already know benefits of adopting new technologies for a better life.

This paper presents a detailed analysis of social benefits of Digitalization of Terrestrial Video Broadcasting Services followed by a comparison of technology formats that are available for selection.

The paper discusses an engineering approach in selecting a technology format that considers the technical aspect of competing for terrestrial TV standards namely DVB-T, ISDB-T, DTMB & ATSC and financial & other pertinent considerations. The paper presents a decision-making frame work that takes into consideration technical aspects, financial & other relevant considerations included in the paper.

Finally paper briefly elaborates the necessity of implementation of Digital Terrestrial TV Network as a sustainable business venture that brings benefits to all stakeholders.

Discovery of things available in the environment such as fire was the first phase of life style improvement cycle soon replaced by seems never ending the second phase of new technological inventions that takes place even now as we read this paper and will continue until the end of civilization. Nations in the world broadly fall into two categories in the technology landscape namely,

Inventors (drivers) – those who develop new technologies and drive standardizations. The inventors or market drivers collectively make every attempt to discourage others joining their club always ensuring as large as a possible market that patronages their new products while individually competing for each other by way of introducing different standards to secure a larger individual share of the market.

Adopters (Followers) – Nations that adopt new technologies developed by others in pure form and or incorporating own limited innovations to improve their lifestyles. Obviously, it is very challenging for adopters to decide what to adopt as they are endlessly facing the blinding glare of campaigns by inventors introducing their new technologies with subtly differentiating features of competing standards aimed at securing a larger share of the market.

So when it comes to adopting new technologies to enhance living conditions of the society by a country; the adoption process must be engineered to achieve the optimum outcome.

2. INTRODUCTION OF TELEVISION BROADCASTING TO SRI LANKA

The Analogue Terrestrial Television Broadcasting Service was introduced to Sri Lanka in April 1979[1]. The television service transformed the society in multiple ways; it changed behavior patterns, a way of thinking and influencing social opinions more than other modes of media, generated new employments & new businesses. Overall the introduction of Television broadcasting, as a technology, is accepted to be beneficial to the society, while negatives mainly due to inefficiencies or inability to change in the face of new challenges by the negatively impacted entities and or deliberate misuse of Television Services by certain elements of the society.

Analogue TV & Digital TV Explained

Analog TV

From broadcasting technology point of view, the technology currently being used in Sri Lanka for television is analog technology. Broadcasters using analog television systems use one of NTSC, PAL or SECAM color system which contains the all necessary information to extract the sequence of pictures by the television receiver, and then use RF modulation to modulate this signal onto a Very high frequency (VHF) or Ultra high frequency (UHF) carrier dedicated to television channel so that it can be transmitted from the broadcaster's studio to the viewer's location. Each frame of a television image is composed of lines drawn on the screen. The lines are of varying brightness; the whole set of lines is drawn quickly enough that the human eye perceives it as one image. The next sequential frame is displayed at a certain frame rate, allowing the depiction of motion to the human eye.

Digital TV - The picture sequence and corresponding audio are digitally processed and converted in to a bit stream (Analog to Digital conversion), which then multiplexed with similar bit streams of multiple channels and then digitally modulated on to an Ultra High Frequency (UHF)

carrier that will carry the bit stream to the receiver. The television receiver reconstructs the picture sequence by using received bit stream (Digital to Analog conversion).

It is important to note that ,in analog version, the receiver extract the picture from the received signal where as in digital version the receiver reconstruct the picture by using the bit stream received.

3. SOCIAL BENEFITS OF DIGITAL TERRESTRIAL TELEVISION BROADCASTING

Enhancements and improvements to the Society are fundamental motivational factors for engineers to develop new technologies. Once developed, the society always embrace new technologies that bring benefits and Digital Television Broadcasting is no different to others in that respect which is well documented and accepted for. Hundreds of countries worldwide already adopted Digital Terrestrial TV Technology. However for the benefits of the readers, this section briefly present key benefits to each stakeholder category [2].

General Public

The Digital Terrestrial Television Broadcasting Networks are generally planned to accommodate all Free to Air channels and many more with a set quality of services matrix across the county. Which means, provided Sri Lanka follows the same general practice with appropriate regulatory and policy framework, general public will have the opportunity to view all free to air channels with the same service quality with minimum or no investment (except for set –top –box) at the receiver end if the network is developed to provide indoor coverage for at least 95% of the household in Sri Lanka; or gone are the days having multiple Yagi antennae mounted in 30 – 40 ft high poles in the house as we can see widely in suburbs & rural areas of the country. Unfortunately, in some areas the viewers are currently forced to subscribed to cable TV networks to view free to air channels as existing analog TV signal reception is so weak to a level that the receiver cannot produce viewable pictures.

Wider choice TV services, improved picture and sound quality, potential portable and or mobile reception (provided regulatory framework in place), electronic programming guide & Enhanced Teletext services with graphics, potential interactive services with the support of Hybrid Broadcasting platforms (HBB), High resolution TV channels such as High Definition (HDTV) & in future upgrades 4K Ultra High Definitions (UDHTV), and broadcasting services for differently-abled persons are among the key benefits that general public as television viewer are going to have. It must be noted that some of the benefits are linked to the regulatory framework and commercial and institutional policies in the countries, which may have to be appropriately amended to enable such benefits reaching the population. It is important to set a road map as a part of terrestrial TV digitalization process to realize all benefits stated here but not limited to.

Broadcasters

Each broadcaster need to have own (or time-shared by multiple parties) nationwide broadcasting network in the case of analog broadcasting, which acts as one of the main entry barriers for new startups limiting viewers a wider choice. Also, as we experience currently in Sri Lanka, not all free to air TV channels are viewable to all households of the country presumably due to deliberate limited coverage of the broadcasters predominantly as a result of cost constraints in extending network coverage to rural areas that may not have much commercial returns in their estimates and or assumptions; a social discrimination continuing since inception of terrestrial TV services in Sri Lanka. Digital Terrestrial TV Broadcasting offers significantly lower capital cost and lower per channel operating cost that will act as a catalyst to start new free to air channels of general contents as well as specialized contents such as education, science, history, etc.

Increased market competition and innovations in content creation as a result of the potential arrival of new entrants and easier storage/processing of content are also among main advantages to the broadcasters.

Whole new portfolio of services and channels that are going to be available is going to attract new viewers and viewing duration of existing viewers expand the market size of the broadcasters.

However, it must be noted that existing analog TV broadcasters will have a significant financial cost

in terms technological obsolescence of existing network assets due to switching off analog TV broadcasting networks that would hit their profit and loss account, which would not be cushioned by any immediate increase in revenue as a result of moving into digital platform; sometimes they might come across reduction in revenue due to increased completion by new entrants. Unless the regulators and policy makers pay due attention to these problems and provide appropriate reliefs for all or those who would need such support to sail through the storm, they would resort to being disruptive towards digitalization process.

Industry Benefits

The introduction of Digital Terrestrial TV to a country is going to bring whole lot of benefits too many industries across. Firstly, the broadcasting industry itself getting widened as a result of new entrants thereby increased productivity and creativity due to higher level of competition, introduction of new services such as pay TV, store and retrieval of programs, real time interactive TV as against current unicast or limited interactive nature and new services. Total revenue of the industry as well as profitability can be increased substantially provided that the implementation is done based on sound regulatory and policy framework. There are many countries in the world those who have digitalized TV broadcasting especially with multiple Digital Technology Standards ending up on the losing side or in a chaotic situation in comparison to analog broadcasting they had before; to begin with, viewers are required having multiple sets – top-boxes in the home. Hence sound and insightful regulatory and policy framework are a critical need for a success story.

Expansion in content producing industries caused by increased demand for content as a result of new channels resulting in the creation of more jobs on the services side.

Digital Dividend is the key benefit a country is going to have by migrating completely to a Digital Terrestrial TV Broadcasting system. Radio Frequency (RF) bandwidth, which carries a single analog TV service channel, can carry multiple of Standard Definition (SD) Digital TV service channels in the range of 4 – 12 depending on the video compression technology adopted. Which means that when a country migrate to a digital Terrestrial TV Broadcasting network completely, the new digital network needs only a fraction of total radio frequency bandwidth occupied by total number of analog TV broadcasting channels setting free a section of radio frequency spectrum used before, known as the Digital Dividend, that can be used to expand broadcasting itself to introduce more TV channels, higher quality TV channels(HDTV, UHDTV) or more importantly for mobile broadband such as 3G & 4G in order to reduce the digital divide by providing broadband internet access to geographical areas which are still not covered by 3G or 4G or to reduce congestion in built up areas.

4. DIGITAL TERRESTRIAL TV BROADCASTING TECHNOLOGY STANDARDS

Migrating to digital terrestrial TV system is going to benefit everybody in a country; therefore, the mood is let us do it quickly.

However, similar in selecting a color TV System (NTSC, PAL, SECAM, etc) in analog TV broadcasting, the adopting countries are faced with the challenge of adopting a Digital Terrestrial TV broadcasting technology standard. Table 1 presents Digital Terrestrial TV standards available for selection.

Table 1: Digital Terrestrial TV Standards

Standard	Promoters	Implementations
DVB-T2*	Europe	Europe & Asia
ISDB-T**	Japan	Japan, South American &
ATSC	USA	North America &
DTMB	China	China & Asia

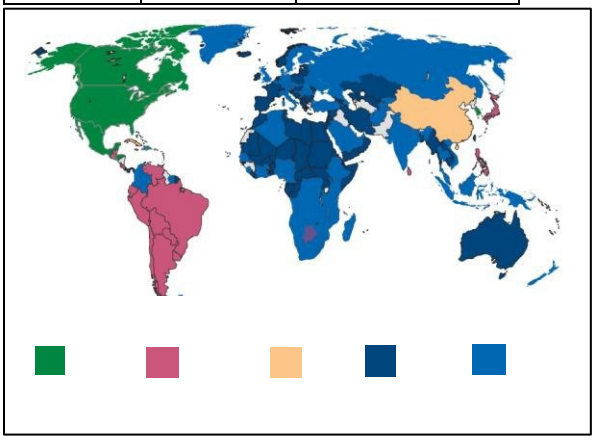


Figure 1 depicts the worldwide adoption/implementation of Digital Terrestrial TV standards [3].

Source: www.dvb.org

Figure 1: Worldwide adoption/implementation of Digital TV

5. ENGINEER DEPLOYMENT OF DIGITAL TERRESTRIAL TV BROADCASTING NETWORK

The Oxford English Dictionary[4] defines the word "Engineer " *A skillful contriver or originator of something* as a noun and *Skilfully arrange for (something) to occur* as a verb; definitions, among many others in the dictionary, that would be the best fit for this case.

So engineer deployment of Digital Terrestrial TV broadcasting network needs consummately meticulous evaluation of most important if not all, parameters relevant to a country to come up with the best-fit choice to live up to the definition of the word "engineer".

	DVB-T2	ISDB-T	DTMB	ATSC
RF Transmission	COFDM (6,7,8 MHZ)	COFDM (6,8 MHz)	SC or TDS-OFDM (2/8 MHz)	SC
Modulation	QPSK, 16, 6	DQPSK,QPSK,	QPSK, 16,32,64	8VSB, 16VSB
FEC	BCH+LDP	RS+CC	RS+LDPC	RS+CC
V Codec	MPEG-2, H.264/MPEG4AV C	MPEG-2,	MPEG-2, H.264/MPEG4AV C	MPEG-2, H.264/MPEG4AV C

Factual necessity of investing in a Digital Terrestrial TV Broadcasting network among other important priorities that demand allocation of limited resources, preparedness of the society to accept and enjoy such a service, cost of technology at the time of the decision as the cost of new technology always reduces along the time axis, the ability of households to invest on

This paper focuses on the part of the engineering process that comes after the decision to deploy a Digital Terrestrial TV Broadcasting Network, which needs to consider two main parameters namely;

Technical aspects of competing Digital Terrestrial Broadcasting Standards

Financial & other considerations

5.1 Technical aspects of competing Digital Terrestrial Broadcasting Standards

Multiple countries started development of technologies and platforms simultaneously to deliver digital terrestrial TV services; as a result of which there are multiple standards in the world competing each other for market share. Those competing standards, while intensely promoting each one superior than others, are upgrading their capabilities to better levels (new versions or upgrades) in quick succession along the time axis making what is best technically today as a selection not to be the best digital terrestrial TV standard technically at the time of implementation, which naturally takes around 2- 3 years of time to complete at earliest.

Table 2: Selected Technical Aspects of Digital Terrestrial TV Standards

Notes:

** DVB-T was the early version of DVB-T2 and DVB-T & T2 together claims the widest adoption regarding geographies and population.*

*** ISDB-T had multiple versions and grouped together.*

Abbreviations

COFDM – Coded Orthogonal Frequency Division Multiplexing

RS – Reed-Solomon SC Single Carrier

CC – Convolutional Code

LDPC – Low-Density Parity Check

Therefore, the most practical and widely adopted engineering choice by adopting countries is to take a snap shot of technical capabilities of all available digital terrestrial TV technology standards and combine with other important non-technological evaluation criteria such as financial, social, political etc under consideration so that finally to come up with overall most beneficial digital terrestrial TV standard. Table 2 depicts such a snapshot gathered from information available in the official websites maintained by respective TV standards and in the public domain. It is important to note that the year indicated in each column.

Japan launched in the year 2007 a new version of ISDB-T standard called ISDB-T International that has enhanced capabilities than ISDB-T first version launched in the year 1997[5].

ATSC also has announced that they are working on ATSC 3.0 standard that will have many advanced technical capabilities that the current Digital Terrestrial TV standards offer[6].

RF transmission, Forward Error Correction (FEC) and modulation type are directly linked to the spectral efficiency of the RF channels measured in bits/s/Hz that determines the maximum capacity in terms of aggregate bit rate (Mbps) of a channel for a given bandwidth.

Standard			I	FFT	GI
ISDB - T	Layer A	NSEG = 1	4	8k	1/8
	Layer B	NSEG = 12	2		1/8
	Layer A	NSEG = 1	4	8k	1/16
	Layer B	NSEG = 12	2		1/16
DTMB	PN420			4k	1/9
	PN595			SC	1/6
	PN945			4k	1/4
ATSC				SC	
DVB - T2				32k ext	1/128

Standard	FEC		Modulation	BW [MHz]
	inner	outer		
ISDB - T	2/3	188/204	GPSK	6
	3/4	188/204	64-QAM	6
	2/3	188/204	GPSK	6
	3/4	188/204	64-QAM	6
DTMB	0.8		16-QAM	8
	0.8		16-QAM	8
	0.6		16-QAM	8
ATSC			8-VSB	6
DVB - T2	2/3	BCH	256-QAM	8

Standard	Data Rate [Mb/s]		SE [bit/s/Hz]	C/N dB Gaussian
ISDB - T	0.416	17.27	2.88	5.5
	16.85			19.4
	0.441	18.28	3.05	5.5
	17.84			19.4
DTMB		21.66	2.71	12.8
		20.79	2.6	13.1
		14.44	1.8	10.4
ATSC		19.29	3.22	15
DVB - T2		40.21	5.03	17.8

Table 3: Spectral Efficiencies of Digital Terrestrial TV Standards

Support for latest video compression codes defines how many TV programs (Channels) such an RF channel could transmit simultaneously. Table 3 depicts calculated spectral efficiencies of different Digital TV broadcasting standards.

The spectral efficiency and corresponding minimum C/N ration are technical parameters widely used to compare capacities of RF channel bandwidth each digital terrestrial standard can deliver. Highest spectral efficiency at the lowest C/N ratio ideally is the best regarding capacity, but practically that is more desirable than available. It is prudent to select the digital technology standard with higher spectral efficiency even at a higher C/N figure as such selection maximizes the Digital Dividend. However, following aspects are also should be considered.

1. Successful deployment history and total deployments in the world regarding geographies and population.

2. The future development road map.

3. Guarantees offered by equipment suppliers to upgrade to future versions with minimum hardware changes, which eliminates short term technical obsolescence of the hardware deployed in the middle of asset depreciation cycle.

4. Risk of becoming dead technology Standard in the face of competition.

At present DVB-T2 standard seems to be the best technology standard from the technical point of view while others, especially ATSC and ISDB-T are sending positive signals of coming up with technically better versions in future.

As history proved with respect to competing color systems of analog TV such as PAL, NTSC & SECAM, the ultimate survivors will be matured to achieve best technically possible capabilities reaching a natural end to technical superiority based competition by digital terrestrial TV standards that we witness today.

Table 4 depicts evolution timelines of each technical standard.

Family	Version	Launch
DVB-T	DVB - T	Year 2000
	DVB-T2	Year 2008
ATSC	ATSC 1.0	Year 1996
	ATSC 3.0	Not yet
ISDB-T	ISDB-T	Year 1997
	ISDB-T	Year 2007
DTMB	DTMB	Year 2006

Table 4: Evolution time lines of Digital Terrestrial TV Standards

Therefore it is very important to consider financial, commercial and other country-specific parameters in arriving at the final choice of Digital Terrestrial TV standard.

5.2 Financial & Other Considerations

The majority of adopters are resources constrained countries that have gotten many competing priorities demanding allocation of limited resources among which financial resource generally is the scarcest. So it is essential to go diligently through financial & commercial aspects. Following are some of the vital considerations

The cost of the new network - The total cost of new network based on selected Digital Terrestrial TV standard. Naturally technically superior standards demand a price premium over others. If such price premium exists, accommodation of such a premium must be justifiable with respective to immediately realizable revenue arising out of increased Digital Dividend or any other technical aspect considered. Technical standard of higher spectral efficiency provides higher Digital dividend, but if there is no immediate or short-term demand for incremental radio spectrum released, then there is no financial compulsion to select the standard having the best spectral efficiency at a premium price.

The cost of Receivers - Latest high-end TV receivers have built in digital tuners of one or few technology standards, whereas low-end and old units do not, requiring households to invest for set –top boxes. Even though most of the current TV receivers carry digital tuner corresponding to one or few Digital TV standard/s, the price of modern TV receivers does not seem dependent much on standard digital tuners built in. The price of TV Recover is mostly a function of display technology and other user-focused fancy features. The cost of set-top boxes mostly seems in far with each other with respect to all Digital Terrestrial TV Standards. The price of set –top box is mostly driven by minimum order quantity than digital terrestrial TV standard.

The cost of the analog switch off – Analog switch off immediately forces existing analog broadcasters to fully depreciate or write off the value of broadcasting assets that carry in the balance sheet impacting their financial performance and other financial related obligations. As an example, if those broadcasting assets are funded through bank borrowings with assets as collaterals then banks would require immediate settlement of outstanding loans triggering cash flow stresses. Offering the same content through new digital terrestrial TV platform would not increase revenue and enhance profitability overnight. Therefore, a financing scheme to compensate existing analog broadcasters would be necessary, and that cost is also must be part of the digitalization cost. It must be noted that this compensation cost is not Digital Terrestrial Standard dependent.

Compensation for loss of Jobs – Digital Terrestrial TV transmission network would not

require large maintenance work force, therefore there would be loss of jobs especially in network maintenance area if there are too many analog broadcasters in the country having own maintenance departments to operate networks on 24x7 basis; and all of them would not be required to operate the new digital network. Compensating them as per the regulations must be built in the total cost; this cost is also not standard technology dependent.

Attractive Funding schemes – Promoters of some Digital Terrestrial TV standards offer very attractive funding option if the country adopts their technology standard, of which current version would not be the one that offers the best set of technical advantages. The saving on cost of capital and corresponding loss of technology premium appropriately valued in terms of financial terms must be considered & incorporated in to the model.

Social & Political Considerations - Some countries have special social & political relations with promoting countries of certain standards as a result of which adopting country makes an adoption decision due to such social and or political considerations. This is a special situation, not uncommon, though, and out of the scope of this paper as the adoption is not value based.

5.3 Decision Making Framework

Having established critical technical, financial and other parameters that must be considered in deciding most suitable Digital Terrestrial TV Broadcasting standard, now a frame work could be formed as outlined below.

Technical comparison of the available set of Digital Terrestrial TV Broadcasting Standard is fairly straightforward. For a desired C/N value, which defines the coverage profile and capacity regarding Mbps per transmission site to a given transmit relative power reduction in Digital Dividend can be assessed and tabulated in comparison to the technically best digital TV standard at the time of evaluation.

For Overall evaluation purposes reduction

In Digital Dividend estimated at the opportunity cost (price at which freed RF bandwidth could be sold) can be added to each technology standard under consideration as a recurrent cost.

Estimating total network cost for a given C/N ratio, which governs the network availability to viewers is a network planning job that must be done for all technology standards under consideration. Field trials are not necessary if all technology standards, under consideration, have commercial deployments, in which case field data is readily available. Otherwise field trials are essential to establish performance parameters.

Funding schemes by promoters must be carefully evaluated against conditions attached such as

Restricted procurement – funding party sometimes limits procurement freedom that results in the recipient having to pay a premium over the free market price. Such premiums can be estimated using price bench marks and must be included as a cost.

Professional Services – funding party makes it compulsory to use professional services from them up to a certain value irrespective of such expertise available in the recipient country or could be procured at a lower cost from elsewhere. Such cost differentials must be included in the cost.

Private Equity – The country, through an appropriate procurement process, can license a private company/companies to build own and operate the Digital Terrestrial TV network/s as an infrastructure services entity/ies that provide channels to contents broadcasters on a commercial basis. For countries of small geographical size with a small population, it would not be viable to license multiple players to crate competition at the infrastructure level, which may in turn perhaps reduce Digital Dividend. Therefore the most prudent approach would be for such countries to create monopoly and license could be granted to the party that commit the best business model for a given capital commitment. Conversely large counties with large populations having densely populated cities could create competition at infrastructure level by allowing multiple players to operate at the infrastructure level, however allowing multiple technical standards must be coupled with availability of TV receivers and Set-top boxes supporting all adopted multiple technology standards.

This five point frame work would help to establish the total cost of procurement of each technology standard under consideration so that a value based selection decision could be arrived.

6. IMPLEMENTATION OF DIGITAL TERRESTRIAL TV BROADCASTING NETWORK

Traditionally analog broadcasting networks and(Even telecom networks were) developed according to a roll out plan of individual license holders mostly based on a business plan and business strategy of the license holder.

However, Digital Terrestrial TV Broadcasting Network in a country essentially a shared infrastructure network for which media license holders are required to pay for program channels that they are going to use. The payment structure typically would be onetime payment plus monthly/annual recurrent rental. The rental structure would be based on a business model developed by the entity that is entrusted to deploy and operate the Digital Terrestrial TV Broadcasting Network. Therefore the cost model and the business model of the Digital Terrestrial TV Broadcasting Network developer and operator are going to have a critical impact on the business model of analog broadcasters who are compulsorily required to migrate to digital network on a given time schedule (to achieve Digital Dividend) and by new start up broadcasters as the rental structure of channels becomes a cost item beyond their control.

This highlights the critical necessity of value based selection of Digital Terrestrial TV Broadcasting Standard for a country and implementation of Digital Broadcasting Network on a frugal cost structure and a business model agreed by policy makers, regulators and broadcasters.

A proper awareness campaign to educate the general public is also very important especially on the importance of digitalization of terrestrial TV broadcasting; the benefits to the society and make them aware the Digital Terrestrial TV standard adopted so that they can start looking for compatibility when they make purchase decisions on new TV, ability of continuing with old TV with the use of set –top box must be communicated, and traders can make such TV models and set-top boxes available in their stores. Most important importantly the timelines of Digital Switch over (DSO) and Analog Switch off (ASO) must be publicly announced.

Countries that are migrating to Digital Terrestrial TV Networks most have well developed nationwide telecommunication networks and mobile communication networks, which are technically more complex than broadcasting networks; which means enough expertise is available in the country on new technology deployment, network planning, network roll out and project management, so it is not too difficult to develop a practically achievable program with all relevant risks identified and mitigation strategies built in.

There are more than hundred countries [3] that have successfully deployed Digital Terrestrial TV Broadcasting Networks of different technology standards, which means that countries planning to deploy such network have ample case studies to follow and make their implementation plans near perfect if not perfect.

Sri Lanka been in the forefront of adopting net technologies especially in the telecommunication industry and enjoying the benefits of them seemed taken more time to decide a standard for Digital Terrestrial TV Broadcasting. Recently media [7] reported that Sri Lanka decided to adopt ISDB-T standard for digitalization of terrestrial TV, which is a welcome news as it is the most important first milestone in the terrestrial TV digitalization roadmap. Sri Lanka is blessed with huge knowledge and talent base on modern technology deployments on successful business models could look forward to another success story within the relatively short time span.

7. REFERENCES

[1]. Indrajit Banerjee, Stephen Logan, ed. (2008). Asian Communication Handbook 2008. Singapore: Asian Media Information and Communication Centre, Wee Kim Wee School of Communication and Information, Nanyang Technological University. pp. 446–459. ISBN 9789814136105.

[2]. International Elecommunication Union – Radiocommunication Sector Itu-R Faq On The Digital Dividend And The Digital Switchover; pp 1-

13,(http://www.itu.int/en/ITU-R/Documents/ITU-R-FAQ-DD-DSO.pdf)

[3]. www.dvb.org

[4]. http://www.oxforddictionaries.com/definition/english/engineer

[5]. http://www.dibeg.org/

[6]. http://atsc.org/newsletter/atsc-3-0- where-we-stand/

[7] .http://www.dailynews.lk/?q=2016/05/30/local/83128

Power Analysis Based Side Channel Attacks

Hasindu Gamaarachchi[1], *Harsha Ganegod*[2], *Roshan Ragel*[3]
Department of Computer Engineering, Faculty of Engineering, University of Peradeniya, Peradeniya, Sri Lanka
[1]hasindu2008@gmail.com, [2]hbbcool.bhanuka@gmail.com, [3]roshanr@ce.pdn.ac.lk

ABSTRACT

Power Analysis Attacks which break the key of a cryptosystem by measuring the power consumption of a device have become a huge security threat. Advanced Encryption Standard (AES) which takes billions of years to break via a brute force attack can be broken in few minutes using a power analysis attack. Therefore, to minimize the threat imposed, research on countermeasures has become extremely important. First, we build a testbed for power analysis, which is a set of complete hardware and software components that can be used to do a practical demonstration of a power analysis attack. Then using the testbed, we show that even the latest encryption algorithm like Speck can be still broken in a time less than 1 hour. Despite being an add-rotate-xor cipher that does not even use substitution box operations, we show that Speck can be broken not only on an 8 bit microcontroller but also on a 16 bit microcontroller. Next we practically test and evaluate the effectiveness of some selected countermeasures. We work on both circuit level hardware countermeasures as well as software countermeasures. While experimenting with new ideas for countermeasures finally we provide improvements to an existing countermeasure.

Keywords - **Power Analysis Attack, Correlation Power Analysis, Power Measurement Testbed, Speck, Countermeasures for Power Analysis Attacks**

I. INTRODUCTION

Side channel attacks are a type of cryptographic attacks where unintended channels such as heat, sound, power and electromagnetic radiation is given out by a cryptosystem are used for breaking the secret key. Power Analysis Attacks which fall under side channel attacks use the power consumption of the cryptographic device as the side channel. Currently, electronic components such as microcontrollers and memory are made out of Complementary Metal Oxide Semiconductor (CMOS) circuits. In CMOS circuits, power consumption depends on switching of transistors. Therefore, the power consumption of a computational device depends on the data that is being processed and the type of operation being performed [1]. When power consumption pattern (power traces) of a cryptographic device is captured as shown in Fig. 1, it is possible to derive the secret key of a system. Power traces are obtained for encryption on several plain text samples. Then those power traces are analysed on a computer using techniques such as Simple Power Analysis, Differential Power Analysis (DPA) [2] or Correlation Power Analysis (CPA) [3]. This type of attack has hugely affected the security of embedded devices such as smart card. Even a most used encryption such as Advanced Encryption Standard (AES) can be broken in few minutes. Because of the imposed security risks, research on power analysis attacks and countermeasures has become greatly important.

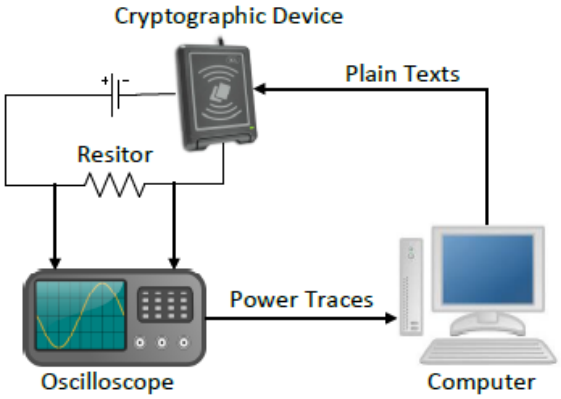


FIGURE 1: Power analysis attack

CPA is a statistical technique which uses Pearson correlation [3]. The power consumption of the device is calculated by using a power model such as hamming weight for each possible key. Then the calculated power data (hypothetical power consumption values) are compared with real power values from power traces using Pearson correlation. The maximum correlation corresponds to the correct key. Due to the advantages CPA has such as requirement of less number of power traces, we use CPA for the research.

The first requirement for power analysis is a test bed. Few works such as [4], [5] and [6] mention the specifications of their testbed but unfortunately no descriptive steps are provided, making it difficult to reproduce. Building a testbed is a complicated and time consuming task especially without a proper guide. We have contributed by providing a step by step guide for building a testbed from the scratch, so that it would save the time of future researchers. Further we have introduced novel power measurement methods that can break AES even with less than 200 power traces in less than 10 minutes time, while just using passive oscilloscope probes.

Most works such as [4], [5] and [6] focus on attacking AES. Speck is a recent light weight cipher introduced by National Security Agency (NSA) that may become famous among embedded devices [7]. This algorithm has lot of differences with AES and hence the attack approach for AES cannot be used. Up to now according to our best knowledge there is no work on attacking Speck. We show that Speck can be broken in less than 1 hour. We contribute by showing the vulnerability in Speck while emphasizing the need for countermeasures.

Various work such as [1], [2] and [8] propose countermeasures against power analysis attacks. But most of them are not practically demonstrated. Some have used simulators to test their countermeasures but in real world, ideal conditions assumed in simulators are no longer there. Therefore, we practically implement some selected countermeasures on the testbed we have created and attack them to analyse effectiveness. Further we check the effectiveness of some ideas of our own while also proposing improvements for existing countermeasures.

II. POWER ANALYSIS TEST BED

A. Making of the Test bed

A power analysis test bed includes hardware components (Fig. 2) such as the cryptosystem and the power measurement setup. Software components include programs that automate the power capturing process and the ones that do analysis.

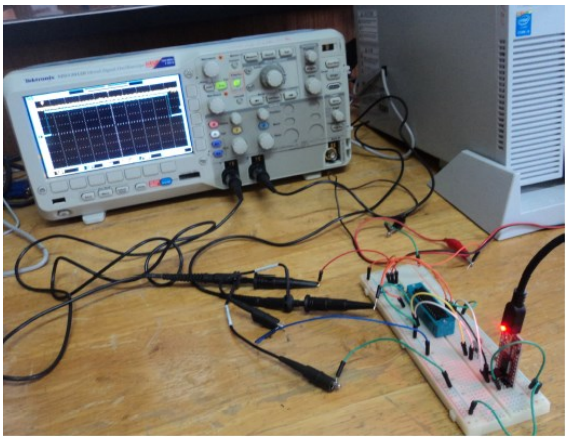


FIGURE 2: The testbed implemented on a breadboard

Fig. 3 shows the circuit diagram of a testbed we have created. The cryptosystem which carries out the encryption is an 8-bit microcontroller. It is programmed to run 128 bit AES. The microcontroller is interfaced with the computer via Universal Serial Bus (USB) using a USB to RS232 TTL (Transistor Transistor Logic) converter shown as FTDI in Fig. 3. Either the Vdd resistor or the Ground resistor marked in Fig. Three can be used for power measurements. The resistor is usually of about 100 ohms. If ground resistor is used, the oscilloscope probe tip is connected to the place marked as Connection Point 1. If Vdd resistor is used, the oscilloscope probe tip is connected to the place marked as Connection point 2. In both cases, the ground of the oscilloscope is connected to the place marked as Ground. The pin marked as Trigger is used to provide the trigger for the oscilloscope.

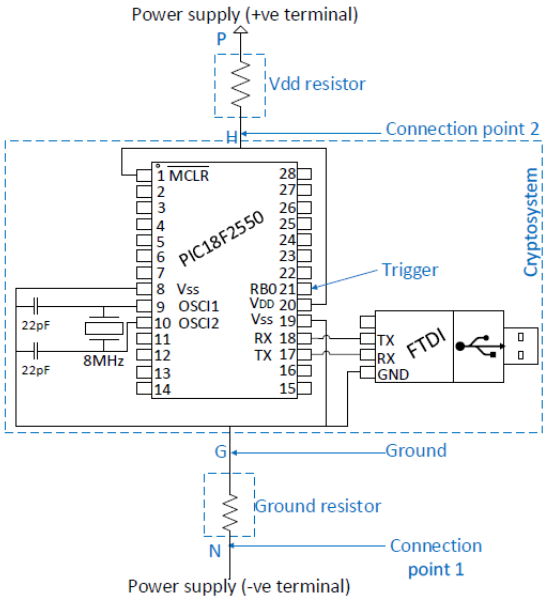


Figure 3: Circuit diagram for the testbed

Works such as [6] when measuring power, measure the voltage across the resistor. Though Vdd resistor method provides better results (less power traces needed for a successful attack) it requires a differential probe which is costly [6]. We introduce a new technique where the voltage across the microcontroller is measured rather than across the resistor that still enables the use of a normal oscilloscope probe when using the Vdd resistor method. Also our testbed even on a breadboard enables a successful attack with only about 200 power traces, despite the high level of noise that would be there when implemented on a breadboard. Further, we have made a working setup that uses the internal USB module of a microcontroller which lets the microcontroller be connected directly to the computer even without a USB to RS232 TTL converter.

A digital oscilloscope interfaced with a computer running Matlab is used to collect power traces. The CPA algorithm which is the analysis part, written in Compute Unified Device Architecture (CUDA) C, runs on the Graphics Processing Unit (GPU). Complete source code repository for the testbed can be found at [9].

B. Results

About 200 power traces were enough to derive the key successfully on the discussed testbed that runs AES. It took only 370 seconds to collect the power traces. The analysis was done on a NVIDIA Tesla C2075 GPU where the time taken was only 8 seconds. Therefore, the total time taken was even lesser than 10 minutes.

Fig. 4 shows how the correlation coefficient changes with the number of power traces used. It contains graphs for all possible keys. Here it is clearly notable that one key lies significantly higher from the others. This one being the highest correlated key turns out to be the correct key. Note that even at 50 traces, the correct key has become significantly higher than the others. Therefore, the minimum number of traces needed to attack is even lesser, which shows that the attack is feasible in very small time.

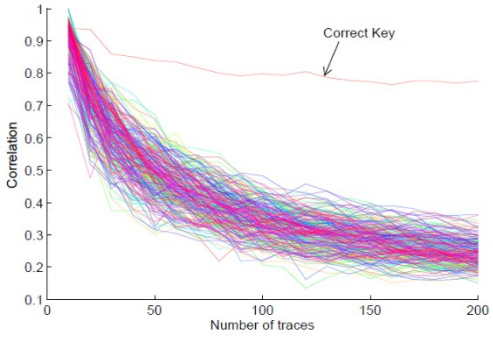


Figure 4: Variation of correlation coefficient with the number of power traces

III. ATTACKING SPECK ENCRYPTION ALGORITHM

A. Speck Implementation

Fig. 5 shows the overview of the Speck algorithm. There it uses only three operations namely Addition, Rotation and Xor (ARX). It does not include any complex operations such as substitution box (Sbox) lookups and matrix multiplication as in AES. Therefore, the algorithm performs brilliantly on low cost microcontrollers.

The reference implementation given by NSA [7] is implemented using 64 bit unsigned integers in C. But due to the lack of 64 bit registers in the microcontroller we used, it was implemented by us for both 8 and 16 bit microcontrollers using small registers.

B. Attack Methodology

The key mixing in Speck happens in a completely different fashion than AES. Also the well-known attack on AES uses the sbox lookup where no such operations are present in Speck. We have introduced methodologies that enable the successful derivation of the key in a Speck cryptosystem.

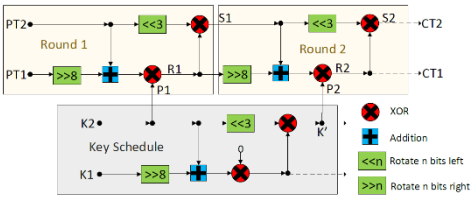


Figure 5: Speck encryption algorithm

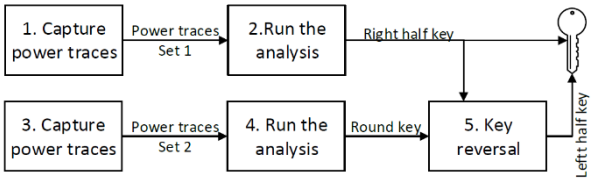


Figure 6: Attack methodology

As shown in Fig. 5 mixing of the key happens in two phases. Right half of the key (K2) is directly mixed at P1 as shown there, but the left half key (K1) is further modified to create a round key (K') that is mixed at P2. Therefore, unlike in AES, the attack on Speck has to go in two phases as shown in Fig. 6. In the first phase, power is collected to include power consumption during memory access of K2 and analysed to break K2. In the second phase, power is collected for K' and analysed to derive K'. If the oscilloscope has good resolution, power can be collected to include both K2 and K'. Then with the help of K2, K' is reversed back to derive K1. Finally, K1 and K2 are concatenated to get the whole key.

Sbox lookup in AES consumes high amount of power. It also facilitates the concept called avalanche effect where the change in one bit of the key completely changes the output. These facts would make CPA easier on AES. As sbox lookups are not present in Speck, the effort needed for the attack increases. The attack methodology introduced by us for Speck, targets the xor operations marked as R1 and R2 in Fig. 5.

Using xor operation which is a very simple operation causes several issues. First due to the lack of avalanche effect, the number of power traces required would increase. Also as xor is not a byte wise operation as sbox, attacking using xor on 16 bit or higher microcontrollers become a challenge. First we addressed the issue by changing the CPA algorithm to attack several bytes at a time. But it is not feasible for 32 bit or higher microcontroller due to time complexity. Therefore, again going back to the byte wise attack, with attack points changed to S1 and S2 in Fig. 5 instead of R1 and R2, a successful attack was realized. Further due to a property of the xor operation (P XoR 0 = P), the key can be always falsely returned as all zeros. This issue was solved by trimming the beginning part of the power traces that caused the unwanted correlation. In depth details of those approaches are not elaborated here as the space is restricted.

C. **Results**

The 8 bit testbed was reprogrammed to carry out the Speck algorithm. The attack which was carried out using 500 power traces became successful. Table I shows the time taken for the attack. As two separate phases are carried out and because more traces are required, time required is higher than AES, but yet it is less than one hour. We also tested on a 16 bit PIC microcontroller, where now the required traces grew up to about 5000. The time required was about 2.5 hours. Therefore, though Speck is a new algorithm yet it is vulnerable to power analysis.

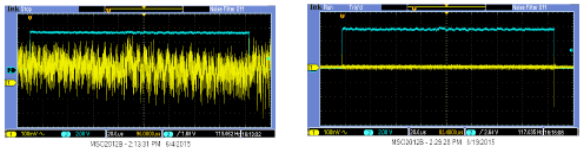
TABLE V : Time Taken for an attack on an 8 bit Speck Cryptosystem

Font Size	Time taken / s
Phase 1: Collecting power traces	913.52
Phase 1: Running CPA	28.97
Phase 2: Collecting power traces	907.63
Phase 2: Running CPA	28.63
Sum	**1878.75**

IV. COUNTERMEASURES AGAINST POWER ANALYSIS

D. Hardware countermeasures

We selected the power line filter based approach proposed by Mangard et al. [1], as no one has practically tested it to our best of our knowledge. A low pass filter is a device that attenuates high frequency components of a signal. The spikes in the power lines when data is read or written, from or to the memory, are the culprits that leak secret information. These sharp peaks are usually formed of higher frequency components. Once a low pass filter is applied to the power line, those peaks are smoothed out, reducing the leakage. Fig. 7 shows how power traces are affected when a filter is used. There, the highly varying waveform is the power trace while the rectangular pulse is the trigger used. When no filter is used, power traces have high peaks and large variations as shown in Fig. 7a. But when a filter is introduced traces become squeezed and flattened as shown in Fig. 7b.



(a) (b)

FIGURE 7: Power traces (a)without filter (b)with filter

TABLE II : Results for Different Filters

Method	Approximate number of traces	Approximate time
Without filters	50	5 minutes
Capacitor (1mF) connected in parallel	1500	1.5 hours
Inductor (1mH) connected serially	500	30 minutes
LC (Inductor-capacitor) second order filter	5000	4.5 hours

TABLE III : Results for Different Components

Method	Approximate number of traces	Approximate time
Without countermeasures	50	5 minutes
Voltage regulator	50	5 minutes
Current source	50	5 minutes
Zener diode	300	20 minutes
Operational amplifier (UA741)	4000	3.5 hours

We implemented several filters and connected them to our testbed. All filters were implemented using passive electronic components such as resistors, capacitors and inductors. The results of the attacks are summarized in Table II. When no filter is connected, for breaking AES even 50 traces are enough. But Table II shows that when filters are implemented, the required number of power traces are increased. But yet the best one out of the ones we tested can be broken in 4.5 hours. This is not a considerable time for an attacker. Therefore, we infer that though filters increase the effort required for power analysis, it is not good enough to make an attack infeasible.

We tried some of our own ideas as well, to see whether they have any effectiveness as countermeasures. Results for those ideas are shown in Table III. Voltage regulator had no effect. Then a constant current source was used to provide power but yet it was not useful at all. A Zener diode connected in parallel to the device increased the number of traces by a little. Then power was supplied through an operational amplifier and it had somewhat better results. But it could be broken in 3.5 hours. Therefore, circuit methods we tested so far are not good enough.

E. **Software countermeasures**

Random instruction injection is a software based countermeasure introduced by Ambrose et al. [8]. The effectiveness has been tested by them only on a simulator and therefore we decided to test it practically on a real system. The CPA algorithm requires all the power traces to be aligned. That is, power consumption during a certain operation should always be at the same time position in all power traces. When some false instructions are randomly inserted in to the middle of the encryption algorithm, this alignment is broken. The more misaligned the power traces are; the more number of power traces are required for a successful attack.

TABLE IV: Results for Random Instruction Injection

Max number of random instructions injected	Approximate number of traces	Approximate time
0	50	5 minutes
1	200	15 minutes
3	500	30 minutes
7	2000	2 hours
15	40000	45 hours

We modified the AEs implementation in section II, so that instructions are randomly inserted. Table IV shows the results of the attack for different number of random instructions injected. The first column is the maximum number of random instructions injected at runtime. The number of power traces required are quadratically increasing and we can predict that injecting about 100 random instructions would increase the attack time to more than 20 days. Therefore, we infer that random instruction injection is effective when compared to the filter based countermeasures described in section IV-A.

The randomness determines the misalignment of the power traces and hence the security basically depends on that. Pseudo random algorithms used for generating random numbers in software always generate the same sequence for a same seed.

In a computer, time would be a good value for the seed. But unfortunately in microcontrollers, real time clocks are generally not there. Currently we are working on a solution where the seed is to be generated using amplified noise signals which are considered to be true random.

V. CONCLUSION

Building a testbed for power analysis attack is the first step in power analysis based research. It is a time consuming task that requires patient debugging especially when a proper step by step guide is lacking. With the testbed we built, AES could be broken in even lesser than 10 minutes time. Speck algorithm has several differences that makes it impossible to use the power analysis approach for AES. But using novel approaches speck could be broken in less than an hour. Though the effort taken to break speck is larger than AES, still it is very vulnerable. Therefore, even for a new algorithm, countermeasures are required. Circuit based countermeasures such as adding power filters and other electronic components were not much effective. But on the other hand software based countermeasures such as random instruction injection proved to be effective.

VI. ACKNOWLEDGEMENTS

The authors would like to thank the NVIDIA Corporation on behalf of the NVIDIA Research Center of University of Peradeniya for sponsoring the graphic cards we used in this research.

VII. REFERENCES

[1] *S. Mangard et al.,* Power Analysis Attacks: Revealing the Secrets of Smart Cards (Advances in Information Security). *Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2007.*

[2] P. Kocher, J. Jaffe, and B. Jun, "Differential power analysis," in *Advances in CryptologyCRYPTO99*. Springer, 1999, pp. 388-397.

[3] E. Brier *et al.,* "Correlation power analysis with a leakage model," in *Cryptographic Hardware and Embedded Systems-CHES 2004*. Springer, 2004, pp. 16-29.

[4] F. Tepanek *et al.,* "Differential power analysis under constrained budget: Low cost education of hackers," in *2013 Euromicro Conference on Digital System Design (DSD)*. IEEE, 2013, pp. 645-648.

[5] Z. Martinasek *et al.,* "General scheme of differential power analysis," in *2013 36th International Conference on Telecommunications and Signal Processing (TSP)*. IEEE, 2013, pp. 358-362.

[6] M. Petrvalsky *et al.,* "Differential power analysis of advanced encryption standard on accelerated 8051 processor," in *Radioelektronika (RADIOELEKTRONIKA), 2013 23rd International Conference*. IEEE, 2013, pp. 334-339.

[7] R. Beaulieu *et al.,* "The simon and speck families of lightweight block ciphers," 2013.

[8] J. A. Ambrose, R. G. Ragel, and S. Parameswaran, "Rijid: random code injection to mask power analysis based side channel attacks," in *Proceedings of the 44th annual Design Automation Conference*. ACM, 2007, pp. 489-492.

[9] H. Gamaarachchi. (2015, Aug.) Poweranalysis. [Online]. Available: https://github.com/hasindu2008/PowerAnalysis

Fuzzy Logic Based Motion Controller For Underwater Remotely Operated Vehicle

P.M. Saranguhewa[1] S.D.S. Lakmal[2], R.M.T. Lakmali[3], N.D. Jayasundere, S.H.K.K. Gunawickrama

Department of Electrical & Information Engineering, Faculty of Engineering, University of Ruhuna
Hapugala, Galle, Sri Lanka

[1]shpmadushanka@gmail.com, [2]sdsameeralakmal@gmail.com, [3]rmtlakma@gmail.com

ABSTRACT

This work demonstrates the applicability of fuzzy controller to an Underwater Remote Operated Vehicle (ROV) for motion controlling with sensor feedbacks. Stabilizing Yaw, Pitch, Roll and depth against external disturbances are considered whilst responding to the remote manoeuvring commands for forward, reverse, lateral and vertical movements. All functionalities are handled by a unified fuzzy controller. Stability and responsivity of the ROV is fine-tuned by adjusting the control parameters. Performance is evaluated using field experiments. Results show the effectiveness of fuzzy controlling of the ROV motion against external disturbances.

Keywords **– fuzzy controller, ROV, control theory, stabilization against disturbances, unified motion control**

1.INTRODUCTION

Remotely operated underwater vehicle (termed ROV) is an underwater robot or an underwater drone. It is a safe and widely used type of underwater vehicle serving a range of military, commercial, and scientific applications. A prototype of underwater remotely operated vehicle was developed with envisaged application of inspecting subsea environment such as coral reefs [1]. As ocean has disturbances like tides, waves and currents, a suitable motion controlling mechanism is necessary for smooth operation of the ROV. The ROV has five propellers with each propeller having more than one type of movement imparted to the structure. Both stabilizing of the structure and maneuvering of the vehicle is done using the same set of propellers. For motion controlling, a unified controller is introduced which takes the feedbacks from the parameter checking sensors as well as the control commands from the operation and gives control signals to the actuators for both maneuvering and stabilization of the structure at the same time. Creating a suitable controller is made difficult by factors like non-linear dynamics of the ocean environment, presence of disturbances as well as observational noises.

There are several controlling strategies that has been developed for the purpose of motion controlling of ROVs. Among them are, supervisory control [2], neural network control [3], self-turning control [4], LQG/LTR (Linear Quadratic Gaussian with Loop Transfer Recovery) [5] and sliding mode control [6]. Recent interests have been in using fuzzy controllers which are known to be effective robust controllers for a variety of applications. For the fuzzy controller, the dynamics of the system is not needed to be fully known. Thus this becomes a suitable candidate for the purpose of ROV motion control. This paper describes the applicability of fuzzy controller for motion controlling, its implementation and results.

2.BACKGROUND

This ROV was developed in our Department in 2013 [1] as a low cost solution for underwater inspection. Integrating improved functionality to this prototype will result in an economical, compact and application oriented ROV. Developed ROV prototype is shown in FIGURE 1. Five motors used as the propellers are shown with labels.

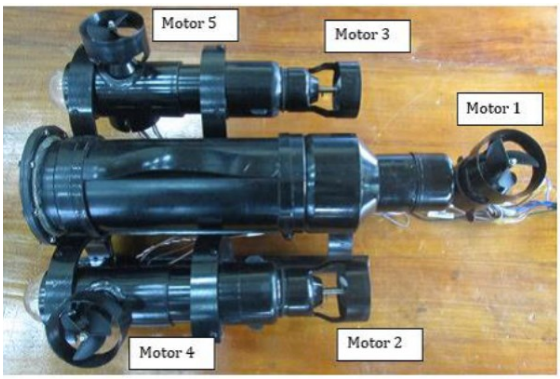


FIGURE 1: ROV prototype with the motors labelled

The main components of the ROV are:

- Five motors with propellers (two horizontal and three vertical)
- High Definition camera
- Accelerometer, Gyroscope, Magnetometer

- Depth sensor
- Ethernet interface (For remote controlling and receiving visual feed & other sensory data)
- Water detector, Temperature sensor
- LED front lights

These components are interconnected using Raspberry Pi, Arduino and network switch modules. Communication methods used are I^2C (Inter – Integrated Circuit), Serial and Ethernet. ROV is connected to a Remote PC using a tether via the Ethernet interface. A client application in the remote PC connects with the ROV. It displays real time data from the sensors of ROV including the live visual feed. Remote controlling of the vehicle is also done using the same application. Either the PC interface or a joystick connected to the PC can be used to control the ROV.

This ROV was designed to be used for subsea inspection. So the medium in which this ROV moves is a dynamic, disturbances prone environment. Ocean contains ripples, waves and currents of various strengths and direction leaving a highly dynamic medium for the ROV. For proper manoeuvring under this environment and keeping the stability of the ROV, good motion control mechanism has to be used. This control system has to reorient the structure against the disturbances and move the structure according to the command signals.

3. CORRECTION PARAMETERS

Considered correction parameters for the ROV are yaw, pitch, roll and depth. ROV has to have movements in forward, backward, lateral and vertical directions according to remote user's command. All these motions are achieved by varying the motor speeds using PWM (pulse width modulation) signals. Respective motors and their directions for each motion is tabulated in TABLE VI.

TABLE 1

MOTOR CONTROLLING DIRECTIONS FOR RESPECTIVE MOTIONS

Motion	Motors	Direction
Forward	2,3	2 and 3 Anticlockwise
Backward	2,3	2 and 3 Clockwise
Upward	1,4,5	1,4 and 5 Clockwise
Downward	1,4,5	1,4 and 5 Anticlockwise
Turn left	2,3	2-Clockwise and 3-Anticlockwise
Turn right	2,3	2-Anticlockwise and 3-Clockwise
Climb	1,4,5	1-Anticlockwise and 4,5-Clockwise
Dive	1,4,5	1-Clockwise and 4,5-Anticlockwise
Roll left	4,5	4-Anticlockwise and 5-Clockwise
Roll right	4,5	4-Clockwise and 5-Anticlockwise

The Table explains the combinations of motors and motor directions required for each motion. In the real scenario, several of these motions has to be applied simultaneously for both error correction and manoeuvring. That makes the number of motor state possibilities higher and each motor has to contribute partially for each simultaneous motion it is involved with. The controller has to respond to both the error amounts measured by the sensors as well as to the remote control commands. The controller must also decide the amounts for each motor that has to be utilized for each required motion and create a unified dynamic motion controlling mechanism for the ROV.

For error correction, errors in the orientation of the structure of the ROV is detected using the magnetometer, gyroscope and accelerometer in the inertial measurement unit. Sensor data contains noise limiting the accuracy of the measurements. These has to be filtered minimizing the degradation of responsively.

For this requirement a MIMO (Multi Input Multi Output) fuzzy controller was developed and implemented on the ROV as its motion controller. Parameters of the controller are adjusted by experimentation for better performance. Effectiveness of this controller on the ROV is evaluated by field testing.

4. FUZZY CONTROLLER DESIGN

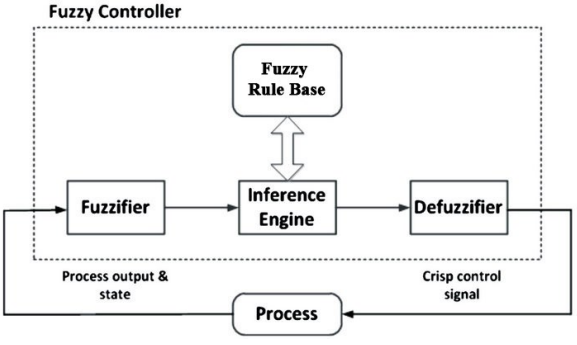


FIGURE 2: Components of a fuzzy controller

A Block diagram of a basic fuzzy controller is shown in FIGURE 2. As illustrated fuzzy controller accepts feedbacks from the process and takes them as inputs. Then it is fuzzified, which is allocating membership values for the inputs values according to predefined fuzzy sets. The fuzzified values are inferred in accordance with the fuzzy rule base, which is also predefined in the controller. The output fuzzy values are then defuzzified in order to get a crisp output value, and a control signal is provided to the process accordingly.

In ROV motion control, the process is the motion of the structure. It has to follow the remote commands of the user and any other undesired movements are taken as errors. Error values are calculated for yaw, pitch, roll and depth in this design and taken as feedbacks from the process to the fuzzy controller. Fuzzy sets, membership functions and rule base is defined according to the requirements. Defuzzified outputs of the fuzzy controller are mapped into the actuators of the design, which are the motors.

Mamdani fuzzy inference method [7] is adapted for this controller. Max-min operators are used for rule inferring. Outputs membership functions are clipped by the evaluated rules rather than scaled. Outputs for each control signal is aggregated and centroid of the resultant is taken as the crisp value for the output.

5. FILTERING SENSOR DATA

Four sensors are used to collect real time data for the fuzzy controller. Data from an accelerometer, magnetometer, gyroscope and pressure sensor is used to calculate yaw, pitch, roll and depth errors. All of this data is scaled to fit into the required range. Offset errors and scale errors were compensated by initial calibration and calculations. Readings from the pressure sensor are amplified for better resolution. All this data contains noise. To reduce the effect of the noise the data is filtered prior to further processing.

Weighted averaging method and Kalman filtering methods were considered for filtering. After testing all methods, the best results were given by one dimensional Kalman filtering for the particular sensor data. A nice feature of this Kalman filter is it can predict future states with the current data. It compensates larger time delays in sensor feedback by predicting intermediate levels.

Following equations are used for one dimensional Kalman filtering of the sensor outputs of ROV.

$$x = x$$

$$p = p + q;$$

$$k = p / (p + r);$$

$$x = x + k * (\text{measurement } x);$$

$$p = (1 - k) * p;$$

The first two formulas represent the prediction of the Kalman Filter. Latter three formulas calculate the measurement update. Variable x is for filtered value, q is for the process noise, r is for the sensor noise, p is for the estimated error and k is for the Kalman Gain. State of the filter is defined by the values of these variables. Those parameters were adjusted to give a clean output without a significant delay. Raw data and its filtered output is shown in FIGURE 3 below.

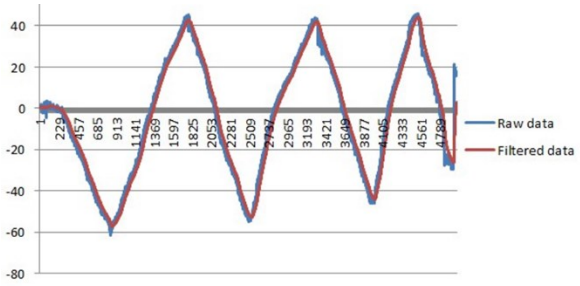


FIGURE 3: Row data and its filtered output

6. FUZZY SUB-CONTROLLERS

There are four main distinct fuzzy controllers required for error correction. They are,

- Yaw controller
- Pitch controller
- Roll controller
- Depth controller

Same abstract controller is adapted for each of these controllers. Controller takes inputs of error and rate of error change of the considered parameter. Set value of the controller is defined according to the initial sensor readings and the remote control values provided by the user thereafter. Membership functions for the fuzzy sets of error and error rate are defined. Also membership functions for the output are defined. A rule base is created with linguistic terms in

accordance with input and output fuzzy sets in this form,

IF (error is negative) AND (error rate is positive) THEN (motor is clockwise); (weight = 0.75)

There are about 10 rules like this for each controller. After inferring the inputs with the rule base, aggregated, defuzzified output is given. Components of the abstract fuzzy controller are shown in FIGURE 4.

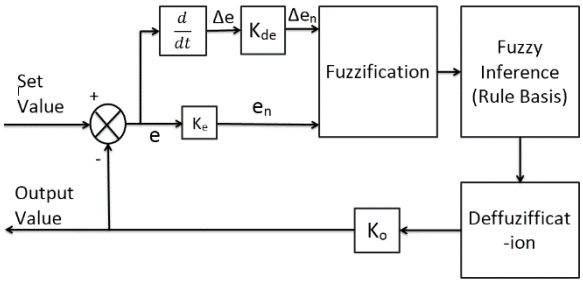


FIGURE 4: Components of the abstract fuzzy controller

If the fuzzy controller for yaw correction is taken as an example, membership functions for the inputs were first selected as triangular functions. With field testing with different shapes, a combination of triangular and Gaussian curves are selected as the optimum membership functions. Selected input membership functions are shown in FIGURE 5 below.

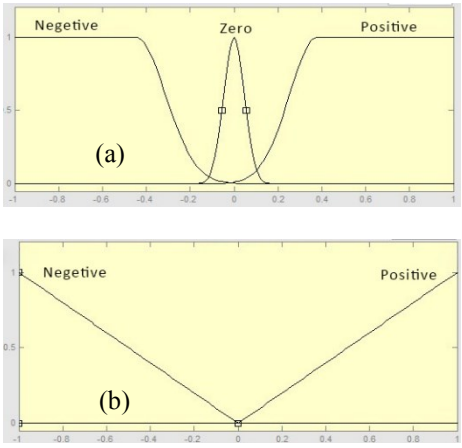


FIGURE 5: (a) Membership Function of Yaw Error, (b) Membership Functions of Yaw Error Rate

Similarly, membership functions for clockwise rotation and anti-clockwise rotation of the outputs are selected. Output of the yaw correction is mapped to motor 2 and motor 3. They are shown in FIGURE 6.

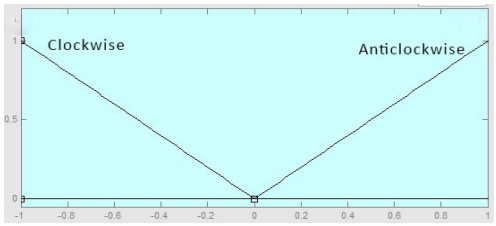


FIGURE 6: Output Membership Functions of Motors

Set value for this yaw correction function is dynamically changed according to the remote commands for left turning and right turning. For pitch, roll and depth similar controllers are constructed according to the respective parameters and motors acting as the actuators. For depth correction error and current vertical acceleration are taken as inputs whereas error and error rate are taken as inputs for pitch and roll.

7. OUTPUT AGGREGATION

In this ROV, several motions are associated with each motor. To achieve unified motion control with both error correction and remote manoeuvring, aggregation of remote commands and fuzzy controller outputs are required. Main output groups mapping to same set of actuators are discussed under following. Components of each group have to be combined in a meaningful manner to give desired motion to the structure.

1) Forward, lateral movements and yaw correction: These three movements are achieved using motor 2 and motor 3 of ROV. Lateral movements (left turns and right turns) are integrated to the yaw correction fuzzy controller as the set value. Combination of forward motion has to be done externally. A fuzzy controller is created to share the available motor power for the required functionality. It gives the amount of power which each function can utilize according to the current yaw of structure, commanded angle to turn and current forward speed.

FIGURE 7 shows how the motor power is distributed for each function. Fuzzy controller deciding the amount to utilize is created with input output fuzzy sets as shown in FIGURE 8 and a rule base written for fair allocation.

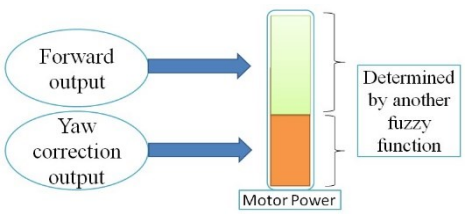


FIGURE 7: Motor power allocation for Forward and Yaw correction

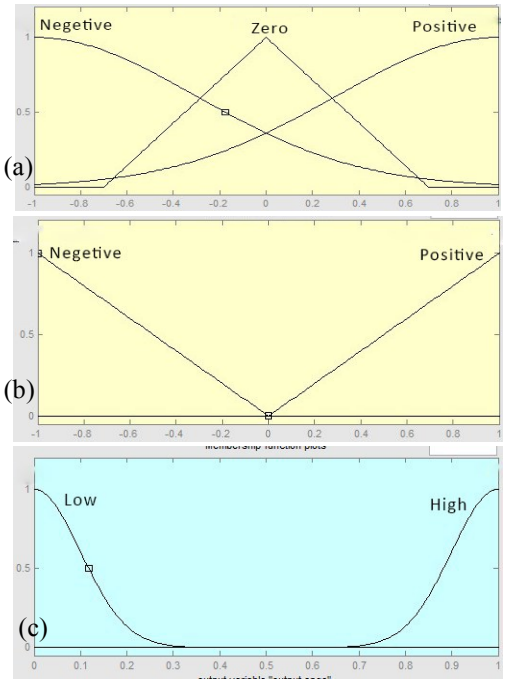


FIGURE 8: (a) Membership Functions for Yaw Error, (b) Membership Functions of Yaw Error Rate, (c) Membership functions for Yaw factor

2) Depth, pitch and roll correction: Motors 1,

And 5 are utilized for each of these movements. As depth is a critical parameter, depth correction is given priority. After allocating motor power to depth correction, rest is distributed for pitch correction and roll correction according to a distribution factor calculated by a separate fuzzy controller as illustrated in FIGURE 9. The Fuzzy controller takes pitch controller output and roll controller output as inputs and calculates the proportion for the two to be allocated for the available motor power. Rule base for this controller is written accordingly. FIGURE 10 shows the input and output fuzzy sets of this controller.

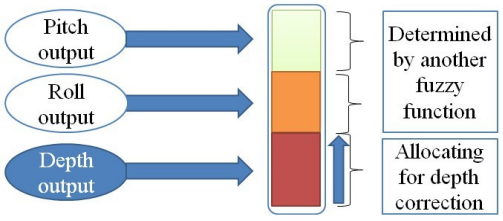


FIGURE 9: Motor power allocation for Pitch, Roll and Depth correction

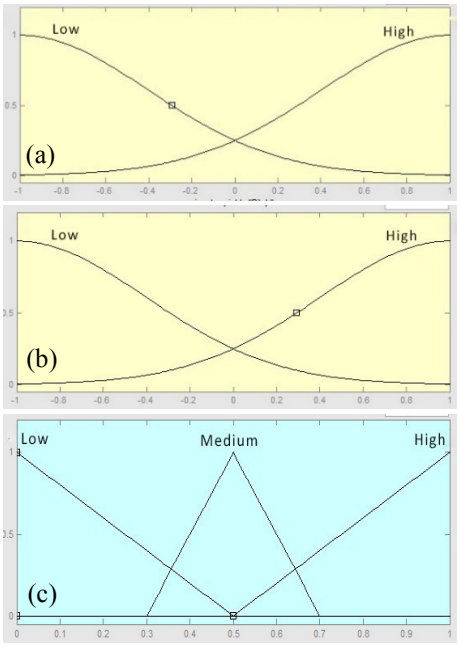


FIGURE 10: (a) Input membership functions for Pitch, (b) Input membership functions for Roll, (c) Output membership functions for the pitch factor

With the combination of all these a fully functional controller is built which takes remote maneuvering commands from user as well as feedbacks from the sensors, take necessary decisions and controls the actuators as a unified fuzzy controller unit.

8. PERFORMANCE

Developed motion controller was first simulated using MatLab. After adjustments it is implemented on ROV using python scripts hosted on the Raspberry Pi module. Computational tools of 'SciPy' python library is used for creating the controller.

It was run without giving disturbances to check if it responds correctly to user commands. Then it is checked for how it responds to disturbances while stationary. Finally, disturbances were given to the structure while it was moving. With field testing fuzzy controller parameters were adjusted for improvements.

Without external forces ROV was responding fast to the manoeuvring signals. Movements were smoother as control commands were also processed by the fuzzy controller in the same manner it responds to a disturbance. When the ROV is commanded to stay still and disturbances were given, it corrected its orientation to its initial state in yaw, pitch, roll and depth as desired. When the ROV is rotated externally at this state, fuzzy

controller corrected the yaw as in the graph shown in FIGURE 11.

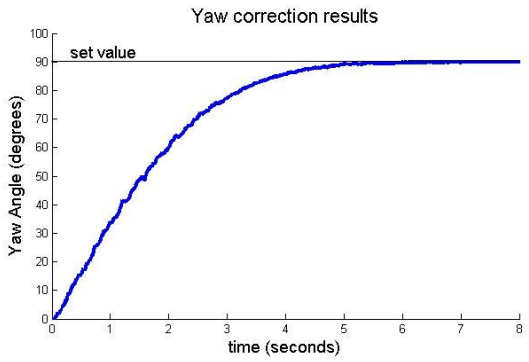


FIGURE 11: Yaw correction results of the fuzzy controller

When disturbances were applied while the ROV was on the move, it adjusts itself with little speed reduction of the original motion at some cases. Every movement including the self-correction ones were relatively smooth, resulting in smooth transitions between motion states of the ROV.

9. CONCLUSIONS

With our fuzzy motion controller, ROV was successfully able to accept and respond to user manoeuvring commands while adjusting it-self when external disturbances are present. This unified controller drove all actuators of the system to achieve the required functionality regardless of the overlapping nature of the control subsystems. Implementing this kind of fuzzy controller would seamlessly improve the performance of a ROV for underwater applications.

10. REFERENCES

[1] Chandimal MP, Chandrasiri HPA, Madusha TPC, Wickramasinghe PGGN, Jayasundere ND, " Underwater Remotely Operated Vehicle", Proceedings Annual Technical Conference 2013 of IET-YP Sri Lanka, 1st March 2013, pp.69-72.

[2] Yoerger, D.R., Newman, J.B., Slotine, J.-J.E., 1986. Supervisory control system for the JASON ROV. IEEE Journal of Oceanic Engineering 11 (3), 392–399

[3] Yuh, J., 1990. A neural net controller for underwater robotic vehicles. IEEE Journal of Oceanic Engineering 15 (3), 161–166

[4] Goheen, K.R., Jefferys, E.R., 1990. Multivariable self-turning autopilots for autonomous underwater vehicles. IEEE Journal of Oceanic Engineering 15 (3), 144–151

[5] Triantafyllou, M.S., Grosenbaugh, M.A., 1991. Robust control for underwater vehicle systems with time delays. IEEE Journal of Oceanic Engineering 16 (1), 146–151.

[6] Christi, R., Papoulias, F.A., Healey, A.J., 1990. Adaptive sliding mode control of autonomous underwater vehicles in the dive plane. IEEE Journal of Oceanic Engineering 15 (3), 152–160.

[7] Ranganath Kothamasu, Samuel H. Huang, 2007, Adaptive Mamdani fuzzy model for condition-based maintenance, Fuzzy Sets and Systems, Volume 158, Issue 24, 2715-27