Permission Based Anomalous Application Detection on Android Smart Phone

Htet Htet Win, Zon Nyein Nway
University of Computer Studies, Yangon
htethtetwin99.hh@gmail.com@*gmail.com,zonnyeinnway7th@gmail.com*

Abstract

Android applications are widely used by millions of users to perform many different activities. Android-based smart phone users can get free applications from Android Application Market. But, these applications were not certified by legitimate organizations and they may contain malware applications that can steal private information from users. The proposed system develops a permission-based malware detection to protect the privacy of android smart phone users. This system monitors various permissions obtained from android applications and analyses them by using a statistical technique called singular value decomposition (SVD) to estimate the correlations of permissions. The training phase emphasizes on the malware samples (approximately 300) downloaded from https://www.kaggle.com/goorax/static-analysis-of-android-malware-of-2017. The proposed system evaluates the risk level (High, Medium, and Low) of Android applications based on the correlation patterns of permissions. The system accuracy is 85% for malware applications and goodware applications.

***Keywords**: Permissions, Android applications, SVD (Singular Value Decomposition), Risk level, Malware, Goodware*

1. Introduction

When we learn the future of mobile, we have found out that there were 2.4 billions of Internet users at 2013. And there will be 5.4 billions at 2025. As the Internet users growth around the world, on the other hand there also increases the people who connect the Internet via mobile. So, we guess there will be 80% of Internet users who make Internet connections via mobile at 2025. Moreover, 50% of transactions will be made by phone at 2050. When we analyze the anatomy of attack, we have found out there are a lot of attacks. They are device attack, network attack or datacenter attack, etc. Moreover, there are many different ways to attack. For example, in device attack, attackers can attack our mobile phone through browser, phone, sms or applications, etc. Among them, our system is intended to detect applications which use unintended permissions. We would like to inform security issues arising from app stores. Attackers can download an application from official app stores and then repackage that application to third party app stores. And as Android Architecture, there are four layers. The first layer is application layer which can interact with end users. And the second layer is application framework. We can develop an android application through this layer.

Android is the powerful operating systems supporting a large number of Application in smart phones. These applications make life more comfortable. With the repaid growing of Android application every day, there are growing threats for the mobile users by installing more malwares without ability to detect them before installing the applications to the user device. Malware name came from “Malicious Software”, its software designed to secretly access a system without the owner’s device knowledge. A key challenge is to identify a suspected application as anomalous (malware). Therefore, we propose the system that can detect the particular app is malicious or not and cancel the installation if the permissions are unacceptable.

The rest of this paper is organized as follows. In section 2, Related Works is presented. The background theory is described in section 3 and proposes system overview is explained in section 4. In section 5, analysis and the experimental results are shown. The conclusion of this system is in section 6.

2. Related Work

Jin Li, Lichao Sun, Qiben Yan, Zhiqiang Li, Witawas Srisa-an and Heng Ye [1] discussed that Scaling the detection for a large bundle of malware apps remains a challenging task. When there are too many attributes of training datasets, decision tree hits a poor accuracy and the training phase tend to take

more time and memory. Instead of extracting and analyzing of all android permissions, develop 3-level of pruning by mining the permission data to identify the most significant permissions. This system achieves over 90% of precision, recall, accuracy, and F-measure using SVM classifier and reduces 4 to 20 analysis times than using all permissions. The limitation of this system is that it cannot grantee advanced obfuscation techniques such as polymorphic and metamorphic malware.

G.N.Bharathi, T.Anusha, R.S.MeenaKumari and P.Aprna [2] presented that most of the users don't understand the permission information and they used to ignore these permission information. At the application installation time, we don't know the apps permission risk and apps may be malicious or not. So, this system assigns a risk score to each app and displays a summary of that information to users. The risk analysis based on the total no: of permissions. The inclusion of risk-score information has significant positive effects in the selection process and can also lead to more curiosity about security related information. The risk calculation is not intended for a user specified application which can be a particular application before installing. Scanning the preinstalled apps are demonstrated as experimental section.

Zarni Aung, Win Zaw[3] proposed that android-based smart phone users can get free applications from Android Application Market. But, these applications were not certified by legitimate organizations and they may contain malware applications that can steal privacy information for users. This system describes the machine learning based malware detection system by using k-mean cluster, decision tree classifier. Both malware and normal applications are classified by using machine learning technique. The limitation is that the system needs to train model with larger dataset to obtain enough samples of malicious applications. And it cannot classify the types of malware application. This system uses a collection of 500 sample android applications and uses Weka tool to analyze the proposed framework.

3. Background Theory

3.1. Mobile Security

When we learn the future of mobile, we have found out that there were 2.4 billions of Internet users at 2013. And there will be 5.4 billions at 2025. As the Internet users growth around the world, on the other hand there also increases the people who connect the Internet via mobile. So, we guess there will be 80% of Internet users who make Internet connections via mobile at 2025. Moreover, 50% of transactions will be made by phone at 2050.

Figure 1. OWASP Mobile Top Ten Risk

Figure 1 shows the mobile top 10 risks which are described by OWASP (Open Web Application Security Project).

When we analyze the anatomy of attack, we have found out there are a lot of attacks. They are device attack, network attack or datacenter attack, etc. Moreover, there are many different ways to attack. For example, in device attack, attackers can attack our mobile phone through browser, phone, sms or applications, etc. Among them, our system is intended to detect applications which use unintended permissions (Misconfigured apps can open doors to attackers by providing unintended permissions).

3.1.1. Security Issues Arising from App Stores

The malicious applications come into our mobile devices from App Stores. Attackers can download an application from official app stores and then repackage that application to third party app stores.

Figure 2. Security Issues arising from Mobile App Store

3.1.2. Android OS

Android is software environment developed by Google for mobile devices that includes an operating system, middleware, and key applications. As Android Architecture, there are four layers. The first layer is application layer which can interact with end users. And the second layer is application framework. We can develop an android application through this layer. The other layers are application framework, libraries and linux kernel.

3.1.3. Malware Detection

Malware detection is a field of study that deals with the analysis, detection and containment of malware. The greatest challenges in security tasks that are still battling the exploration of mobile communication devices, computer and network infrastructures, and web technology are Malware attacks, Malware detection and Malware analysis.

There are three different types of malware detection techniques.

1. **Attack or Invasion Detection:** Tries to detect unauthorized access by outsiders.
2. **Signature-based Detection (Misuse Detection):** Tries to detect misuse by insiders.
3. **Behavior-based Detection (Anomaly Detection):** Detects the patterns in a given dataset that do not conform to an established normal behavior.

3.1.4. Android Permission

The purpose of the permission in android application is to protect the privacy of an Android user. Android applications use a lot of permissions such access the SD card, use the Internet and so on. Google does a great job of declare what permissions an app uses before installing, but does not go into full detail about what exactly those permissions do. When installing an app from the Google Play Store, display dialog box pops up and the user has to agree that the app can do certain things. Every application must have an Android Manifest File. The manifest file provides essential information about the app to the android system.

Figure 3. Sample Android Manifest.xml file

3.2. Singular Value Decomposition (SVD)

The statistical technique, called Singular Value Decomposition (SVD) is the matrix rank reduction technique. It aims to identify the statistical association of permissions. The results of decomposition of the original matrix are descriptions of permissions and apps based on the hidden correlation space derived from SVD. The method reflects the major associative patterns in the dataset.

Consider the M x N Permission_App matrix W.

If W has rank r <= min(M,N), then W can be decomposed by singular value decomposition:

$$W = USV^T \quad (1)$$

Where:

S is also called the “concept matrix”

U is the “permission-concept similarity matrix”

V is the “app-concept similarity matrix”

M is the number of permissions

N is the number of apps

R is the rank value

3.3. Similarity Measure

Another key factor in the success of the proposed system is the similarity measure between testApp and malware apps. There are three simple and well known similarity measures to calculate the similarity. They are the Dice, Jaccard and Cosine Coefficients. Among these three similarity measures, the system is used Jaccard similarity to measure the related permissions patterns in testApp and set of trained malware apps.

3.3.1. Jaccard Similarity Coefficient

The Jaccard index [5], also known as the Jaccard similarity coefficient (by Paul Jaccard), is a statistic used for comparing the similarity and

diversity of sample sets. The Jaccard coefficient measures similarity between sample sets, and is defined as the size of the intersection divided by the size of the union of the sample sets:

Let D = {D1, D2,……Dn) be the collection of N malware apps. Each malware app Di can be represented by a corresponding set Si such that Si is a set of all the permissions contained in Di. Let us denote that set by Di such that Di = {S1,S2,…….. Sn}.

Some attributes are present in just a few objects of a data set. As they assume zero values in most of the cases, they are called asymmetric. Jaccard Similarity Coefficient measure is used to handle asymmetric binary attributes as only non-zero values are relevant for the calculation.

$$S_{A,B}^{Jaccard} = \frac{|\{Per_A\} \bigcap \{Per_B\}|}{|\{Per_A\} + \{Per_B\}| - |\{Per_A\} \bigcap \{Per_B\}|} \quad (2)$$

Where: Per_A = permissions from trained dataset
Per_B = permissions from user-chosen app

4. Framework of the Proposed System

Since the proposed system is implemented by using vector space model, preprocessing stages and indexing are needed. Therefore, query apps and malware apps can be easily and quickly compared. The results are shown in decreasing order of similarity values with the query apps. When we implement the proposed system, we need to pass the following three steps.

4.1. Database Collection

To implement the proposed system, the first thing we should do is to collect the information about the risky apps as much as we can. We need to analyze the nature of risky apps. According to the literature, there are so many ways to analyze different kinds of apps such as by analyzing signature features, behavior features or anomaly features and so on. Among them, we choose to analyze the apps based on permissions. Because permission is the main gate to allow the application (which operations must be done). So, we also need to learn about permissions of android application.

There are a lot of permissions that are declared by Google. Moreover, there are also customized permissions. The specific permission has its own task such as reading contacts, or sending sms or getting GPS, etc. Some of them are dangerous. Some of them are normal. Some of them are nothing meaning etc. But when analyzing permissions, it isn't enough to know which permissions are dangerous and which permissions are normal. One application can use as much as permissions according to the developer. And, we cannot conclude an application has high risk by seeing one of dangerous permission.

So, we need to analyze which correlation patterns of permissions are usually involved in high risk application. To get the correlation patterns of permissions, we choose to apply Singular Value Decomposition (SVD) technique. To apply SVD technique, we need to train dataset to get original matrix (permission-app matrix). For choosing the training dataset, we have to train malware dataset since we give the knowledge that how much risk level has an incoming application. We tried hard to get malware dataset. That kind of dataset didn't download easily as malware based dataset are very restricted. We requested from many sites by using university emails.

After following many agreements and sending many requests to many sites, finally we got the required dataset from https://www.kaggle.com/goorax/static-analysis-of-android-malware-of-2017. Kaggle website describes the specific analysis results of malware applications by separating into four folders. These folders are apkMetaReport, byteCodeReport, virusTotalReport, and assestReport. apkMetaReport folder contains the contents of Manifest.xml files. byteCodeReport folder contains the contents of classes.dex. virusTotalReport folder contains the reports of virusTotal service. assestReport folders contains names of assests and lib contents. So, we choose to download apkMetaReport. That dataset contains over 4000 json files (one json file for one malware application). An android app name is identified by its sha256 hash sum, which is used by file name.

4.2. Preprocessing

We need to preprocess the downloaded dataset to be ready to use as the trained dataset in our proposed system. There are two steps for preprocessing phase: tokenization and removing Duplicate Permissions.

4.2.1. Tokenization

Computers do not understand the structure of a natural language document and cannot automatically

recognize words and sentences. So, humans must program the computer to identify what constitutes and individual or distinct word referred to as a token. Such a program is commonly called a tokenizer or parser or lexer. Tokenizing is the process of breaking of stream of text up into words, phrases, symbols or other meaningful elements.

To store the permissions for each application, we extract the required permissions from the json dataset by tokenization. Then we build original matrix (permission-app matrix).

4.2.2. Removing Duplicate Permissions

Some permission is frequently occurring and that do not represent any content of the application. Duplicate permissions are list of permissions that the developer includes them unintentionally. So, in this phase, duplicate permissions are removed before building the original malware vector.

4.3. Proposed System Design

The proposed system comprises into two phases. The first one is the training phase. The training phase is starting from extracting permissions of each JSON file to performing statistical singular value decomposition (SVD) approach as follow.

Step 1 Place JSON files under 'download' folder of emulator's internal storage.

Step 2 For each JSON file, extract Permissions and do Preprocessing phase. Create a Permission_App relation (perID and appID)

Step 3 Generate the Boolean permission_app matrix and save it to database

Step 4 Compute S, V, U metrics (using Singular Value Decomposition) and reduce the metrics with k dimension (suppose: k=4).

Step 5 Save S^{-1} and U metrics to the corresponding data files

Step 6 Transpose V (malApp vector) and save it to the corresponding data file

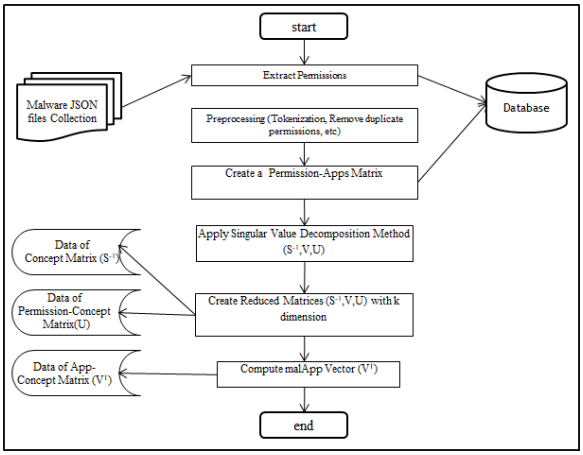

Figure 4. Process Flow Diagram for Training Phase

The second one is the testing phase. That testing phase also contains two sub-phases. First sub-phase is finding the permission-correlation pattern of the user chosen application as follow.

Step 1 Accept a test app.

Get Permission Values of <uses-permission> Element from Android's Manifest.xml (by getRequestedPermissionList() method)

Step 2 Generate testApp vector (q).

Step 3 Calculate queryApp vector by computing q^T U S^{-1}

The second sub-phase is giving message to the user about the user chosen application's risk level as follow.

Step 1 Fetch V^T from corresponding data file.

Step 2 Compute the similarity values between queryApp vector and malApp vectors (V^T) and save them to temporary similarity result array.

Step 3 Choose the highest similarity value (h) from the similarity result array.

Step 4 If h is greater than or equal to maximum threshold value (0.8000), show the message "The application has high risk permissions." in alert box.

Step 5 Else if h is greater than or equal to minimum threshold value (0.5847), show the message "The

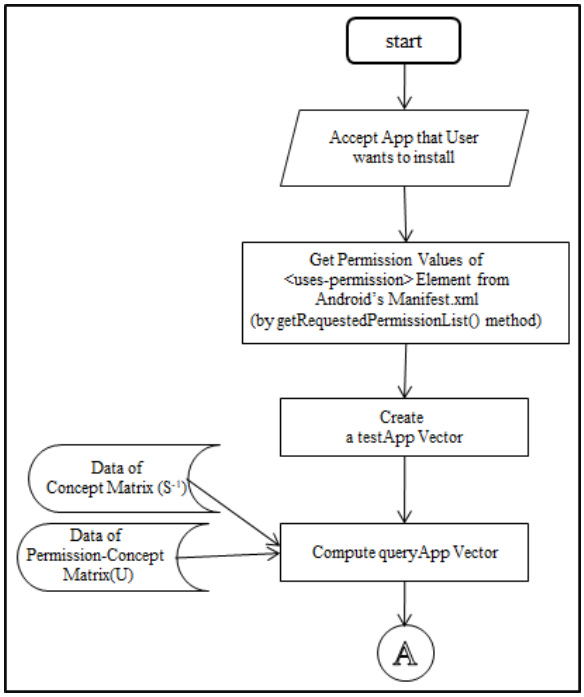

Figure 5. Process Flow Diagram for queryApp Vector

application has medium risky permissions." in alert box. Otherwise, show the message "The application has low risk permissions." in alert box.

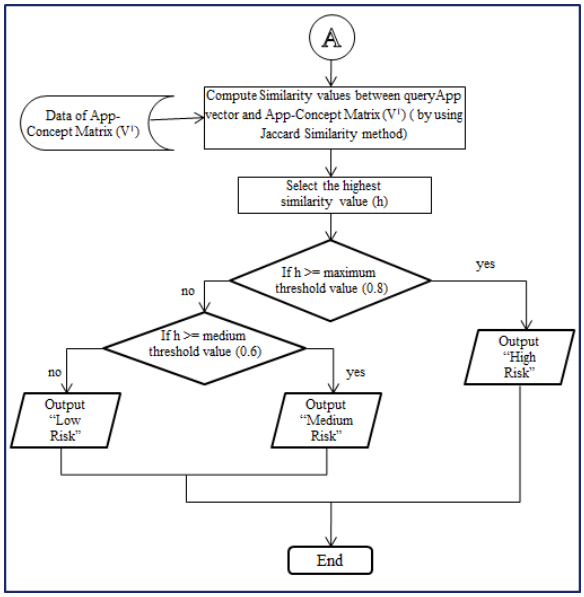

Figure 6. Process Flow Diagram for Finding Risk Level

5. Analysis and Empirical Results

In proposed system, the statistical method SVD and Jaccard Coefficient similarity are used. This system can be used as a malware detection system to decide the incoming app should be installed or not on Android Smartphone.

Firstly, the user wants to install a specific app. That is passed to the pre-processing stages. In database, all significant malware apps are already pre-processing and calculated the malware related patterns. The input app is compared to this malware pattern. Finally, the system displays risky level according to the similarity values.

5.1. Precision and Recall

Precision and recall are two widely used measures for evaluating the quality of results in domains such as Information Retrieval and Statistical classification. In Equation (3) the precision and recall for IR system is shown.

$$precision = \frac{|\{Relevant\} \cap \{Retrieved\}|}{|\{Retrieved\}|} \quad (3)$$

$$\text{Recall} = \frac{|\{Relevant\} \cap \{Retrieved\}|}{|\{Relevant\}|} \quad (4)$$

Precision can be seen as a measure of exactness or fidelity, whereas Recall is a measure of completeness.

5.2. Performance Evaluation

We have over 4000 malware dataset as described in Section 4.1. But we can't train all of that according to phone storage and emulator performance. Firstly we have to find out how much dataset can be trained. So, we train data beginning from 50 dataset. By adding 50 JSON files again and again to the dataset. Here I've found out that emulator was hang at trained dataset 200 on 4G RAM. So, we tried to train dataset beginning from 250 dataset on a laptop which has 8G RAM. We took a lot of time to train that amount of dataset. But we trained dataset by adding 50 JSON files again and again. Here, we have found out that 300 dataset is more suitable on the current situation according to Emulator performance and mobile phone storage.

At that time, we faced another problem that is which 300 dataset will be trained among these 4000 malapps. So, we separated our 4000 dataset by 300. And we trained different 300 datasets on our system to know which dataset has the more features of current environment malapp by testing 95 malapps. Here, we have found out D4 dataset includes the

more features of the current environment malapps. So, we choose that D4 dataset to train on our system.

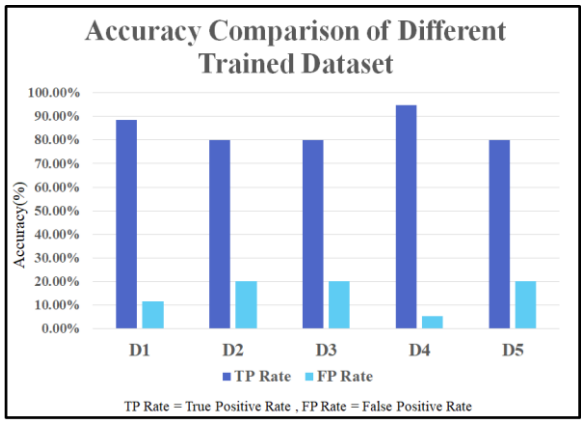

Figure 7. Accuracy Comparison of Different Trained Dataset

The figure 7 shows Accuracy Comparison of Different Trained Dataset. We trained all of 300 separated dataset of 4000 malapps. But at that figure, we just highlight only the appearance dataset.

After getting the best trained dataset, we have to found out which k value will be the best on our trained data according to SVD method. So, we analyze different k value on our trained dataset.

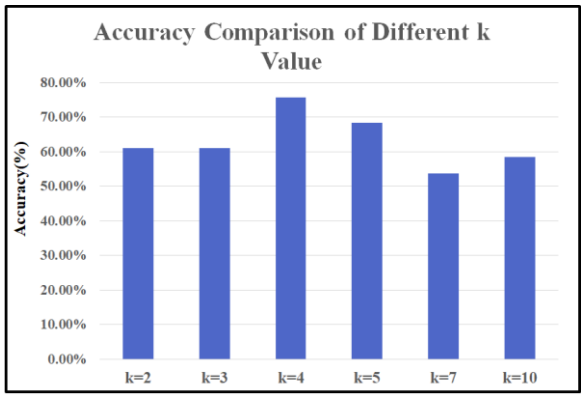

Figure 8. Accuracy Comparison of Different k value

According to the figure 8, we've found out k=4 is the best value of the others. This figure shows the overall accuracy of our system at different k value. But, we highlight just k=2 to k=5 among a lot of different k value.

In figure 9, the correctness of malware is 100% at k=2. But the correctness of goodware is too low. At k=3, also like k=2. At k=4, the correctness of malware decreases a little, but the correctness of goodware is significantly high. So, the accuracy of the system also increases significantly than others. At later k value, the correctness of malware is lower.

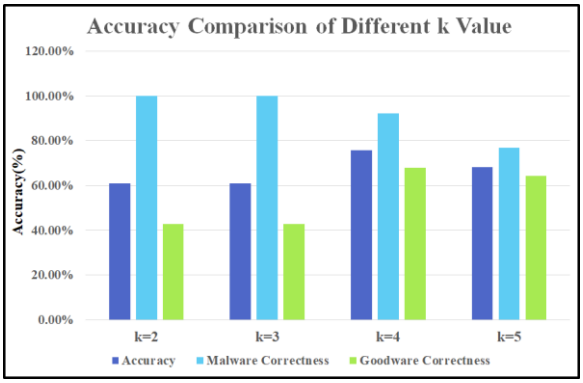

Figure 9. Accuracy Comparison of Malware and Goodware Correctness for Different k Value

So, finally we choose k=4 on the trained dataset 300 according to our analysis of Figure 9 Accuracy Comparison of Malware and Goodware Correctness for Different k Value.

When we analyze the accuracy of our proposed system on 134 applications (96 malwares and 24 goodwares), our system identifies 83 apps as malwares and 13 apps as goodwares on 96 malwares. On the other hand, the proposed system identifies 19 apps as goodwares and 5 apps as malwares on 24 goodwares. So, the overall accuracy of our proposed system is 85%.

6. Conclusion

Focus of attackers and malware writers has changed to mobile devices due to the increased adoption of mobile devices for business and personal purposes and comparatively lesser security controls. Therefore, App stores are common target for attackers to distribute malware and malicious apps.

Our system proposes to detect the risk level for anomalous Android applications. The malware dataset is identified using Singular Value Decomposition (SVD) based approach where a permission-malapp matrix needs to be developed and then query-app can be detected from the set of risky permissions. However, the growing amount and diversity of Android malware has significantly weakened the effectiveness of the conventional defense mechanisms, and thus Android platform often remains unprotected from new and unknown malware.

Our system suggests that the implementation is well suited by finding Jaccard Similarity Values between existing malwares and the user query apps. We can conclude that the Jaccard Similarity measures is well suited for mediate amount of data set and can effectively helpful in finding similar values between

user query apps and malwares. So, the system enables users to search similar risky apps as efficiently and as fast as possible. Therefore, the system can save time in finding the risky apps even we didn't know which apps are closely related and can access the system effectively without an internet connection.

References

[1] "Significant Permission Identification for Machine Learning Based Android Malware Detection", L. Jin, S. Lichao, Y. Qiben, L. Zhiqiang, S.A. Witawas, Y. Heng, 2017.

[2] "Effective Permission Analysis and Complete Security for Android Application", G.N. Bharathi, T. Anusha, R.S. MeenaKumari, P. Aprna.

[3] "Permission-Based Android Malware Detection", Zarni Aung, Win Zaw, March 2013.

[4] "Anomalous Android Application Detection with Latent Semantic Indexing", H. Shahriar, V. Vlincy, 2016 IEEE 40th Annual Computer Software and Applications Conference.

[5] "Ethical hacking and counter measures by EC-Council".

[6] "Spam SMS Detection on Android Smart Phone", Hay Mar Sam, July 2015.

[7] "Information Retrieval for Pharmacy using location – Based Mobile System", Myat Su Aung, December 2016.

[8] "Stock Trend Prediction from Mobile Device Through Web Services", Kay Thi Aung, December 2015.

[9] "SMS Security with Aes Algorithm on Android OS", Phyo Su Khin, February 2017.

[10] "Analyzing Terror Attacks using Latent Semantic Indexing", I. Toure, A. Gangopadjyay, 2013 IEEE.

[11] "A Survey of Android Malware Characteristics and Mitigation Techniques", V. Cooper, H. Shahriar, and H. Haddad, Proc. Of the 11th International Conference on Information Technology: New Generations, IEEE CPS, Las Vegas, USA, April 2014, pp.327-332.

[12] "Using Latent Semantic Analysis to Identify Similarities in Source Code to Support Program Understanding", A. Marcus and J. Maletic, Proc. Of 12th IEEE International Conference on Tools with Artificial Intelligence, November 2000, pp. 46-53.

[13] "Effective Permission Analysis and Complete Security for Android Application", G.N.Bharathi, SSRG International Journal of Computer Science and Engineering – (ICCREST'17) – Special Issue – March 2017, ISSN: 2348 - 8387.