Android App Categorization using Naïve Bayes Classifier

Jagtap A.H.
Computer dept. BSIOTR
JSPM's BSIOTR, Wagholi
Pune, India

Lomte A.C.
Computer dept. BSIOTR
JSPM's BSIOTR, Wagholi
Pune, India

ABSTRACT

This research based on the different integrity protection against various operating based smartphones or cellular phone. A mobile phones, smartphones are essential for daily life but the smartphones based on different operating systems like Symbian, Android mobile devices may be infected malwares because of different applications like Internet app, Bluetooth, MMS, SMS.The malware is a big issue of user mobile security and the downloaded contain through the internet. The security must be provided to smartphone because in advanced the malware after they've been infected they may lose the integrity of our mobile phone. This paper investigates the evaluation and detection of malware through the data mining based technique .The research paper based on Naïve Bayes method, classifier and analysis of the result value by analysis of system. The technique of smartphone content helps to analysis the trusted party. This is mandatory provide the efficiency related to security antivirus and malware detection.

General Terms

Android Security, Malware Detection Algorithm, Precision and call

Keywords

Android, Mobile security, Data Mining

1. INTRODUCTION

The various malware are found due to downloaded application in the smartphones the different viruses, Trojans and spyware also present in the sharing of application via Bluetooth or multimedia message service (MMS), [1] The integrity protection based upon different control mechanism like trusted and untrusted domains .The service provider, user, devices, are the various trusted party whereas the untrusted domain include the downloaded contain via browser, application store in smartphone. Linux based smartphone provides access control mechanisms. The element of these mechanisms is user, User is represented by any integer number or user id and own objects a process or a file, or directory. That can be link to groups. In the Linux Smartphone based file permissions operations each contain of file is divided within an owner user and group IDs and three major contain of read, permission and as well as execute. [2]

Mobile device like smartphone, Linux Based, Cellular Phones have been evolved to be in variety of ways. As it is open source available at anywhere so it is instantly used. It became essential and necessary thing for now a day .The smartphone are user-friendly so security issue arises for manufacturer, provider and user also. The security report [3] Show that different malwares are infected to smartphones and also threat and virus loss the data in smartphone. So that Security functionality is major issue.

2. LITERATURE REVIEW

The review process contain study of existing process, review process include security protection review, malware detection techniques.

2.1 Proactive Security for Mobile Messaging Networks

The mobile phones and other computing devices are used by the users and variety of data transfer through the networks .The data may transfer through the SMS (Short Messaging Service) .The malicious writers put the data related to the messaging networks,[5] this paper proposes proactive security models to protect messaging networks from mobile worms and viruses**.**

2.2 Smart Siren: Virus Detection and Alert for Smart phones

Smart siren is the process which detects virus and alert system for smart phones. The viruses are spread through variety of process like SMS/MMS, Bluetooth, downloaded content and applications

2.3 Android Security

The cellular data plan are rapidly used in marketing, various applications support different services. The enterprise service, social, [6] financial service of mobile smartphone services provide the software installation, marketing .The services helps in accessing various free applications**.**

3. ANDROID MALWARE DETECTION

Smartphone is compact minicomputers, now a day's smartphones plays important role in various applications. It helps to store and retrieve the information, as it is open source platform.[1] There are various operating systems available for smartphone device like Symbian, Android, and Windows phone. In Android malware detection it contains mainly Android based operating system; provide mobile security and malware detection.

The malware behavior found in various application and system. There are various types of malware found in mobile and devices .The dampig, fontal are infected by MMS, Bluetooth, different attribute and services are target to provide protection of malware, only authorization can access the data and malicious code of device. Redbrowser and Mquito. Malwares is by downloading java applications which sends SMS messages to premium rate number at a rate of 30 and 40 rupees per message.

3.1 Android

Android is a Linux based operating system. Now a days android is very popular because it support all services, applications. Android contain different libraries and application program interface which is written in C. Android is kernel based operating system, Linux 2.6, with middleware, libraries and APIs,written in C and application software running on an application framework. Android uses the Dalvik virtual machine with just-in-time compilation to run Dalvik dex-code i.e. .dex file, which is usually translated from Java byte code.

3.2 Mobile Security

The Mobile phones contain various applications, through this application various malware, virus and Trojans are affected to mobile. Thus the security of mobile may break which affect confidential data of mobile user. Security mechanism is based on unauthorized use of mobile application.[4] The security based smartphone contain different configuration and mechanism to give the action to handle the security levels present in the smartphones.

Mobile security provide the integrity protection to devices, the integrity is defined as protection from the untrusted party the integrity in other words is nothing but confidentiality of data i.e. it remain secure in the user authentication only. The protection is done by using many programs and tools used in integrity usage. They protect the damage caused from virus, Trojans and worms and different malwares.

3.3 Data Mining

Data mining is a process of gathering useful information from a large data set. It is process for analysis of data, and finding the content from large data which is called as big data. Data mining is used to summarize the relationships, dimensions and many more terms. Data mining is mining knowledge from data, it extract useful information from huge data .It involves discovery of knowledge ,query analysis and query language, decision tree induction, cluster analysis, and how to mine the Web. Data mining is the process of posing various queries and extracting useful and often previously unknown and unexpected information, patterns, and trends from large quantities of data, generally stored in databases. Data miming contain various technologies which include machine learning, prediction, data management, data warehousing .Data mining developed various applications such as that majorly used in marketing and sales, healthcare, medical, financial, e-commerce, multimedia, and security as well.

The data mining is used in security and its applications, various tools of data mining are available for malware detection. A tool such as detection for email spam, tool for remote exploit detection tool for malicious code detection, tools for botnet detection. Data mining has exploit in cyber security base applications .malware is malicious software which is developed by hacker. Malware includes viruses, worms, Trojan horses, time and logic bombs, Botnets, and spyware. So therefore there are various data mining techniques for detection of malware

3.4 Malware Detection

Malware Detection contain android application file which contain dataset and attributes .The application file send for permission to access the attribute in the application, it is xml file which contain actual object and retrieve object .The use of xml file is that transfer the text file into integer value .the detection of malware uses the mining of attributes, mining refer as future entries. Data mining helps to detect the additional entries in the application .Transform the entries to form one database this entries access using the model (trained model) in which probability of each permission is checked.

The malware detection is based on probability rule, which predict the value and permission access for different applications. It contain data set in terms of matrix which produce output in the form of integer value, which shows the result of detected malware .It is found that malware can be present in SMS,MMS,Bluetooth,Video,Webcam or social networking etc. The malware detection techniques based on Naive Bayes algorithm, which gives approximately correct result.

3.5 Naïve Bayes Classification

For analysis of malware machine learning processes develop one classifier called as Naïve Bayes Classifier, which contain variety of data set. Bayes rule contain probabilistic models. The Bayes rule relies on the statistical properties of a data set and the accuracy of the data set to begin with, so it takes the solution from the statistics as well as data mining. The following formula shows the Naïve Bayes Classifier [8]

The Bayes Naive classifier selects the most likely classification as Vnb be the most likely classification and it will gives different attributes as a1,a2,a3,a4…….an. Thus results in,

$$Vnb = \text{argmax } vj \in V \; P(vj) \, \pi \, P(ai \mid vj) \; . . . . (1)$$

Now elaborate P (ai |vj) as m-estimate which as follows,

$$P(ai \mid vj) = (nc + mp) / (n + m) \; . . . . (2)$$

Where,
n = the number of training examples for which v =vj
nc= the number of training examples for which v=vj and a= ai
p = a priori estimate for P (ai | vj)
m= the equivalent sample size

Bayes rule detect the probabilistic value of variable set permissions histogram, simply make a list including each permission by taking the fraction of the (number of malware having that permission) / total number of malwares in set.

3.6 Mathematical Model

The mathematical model contain dataset value ,precision,recall,android applications ,Permissions ,XML file .The data set values in terms of matrix which contain any values .The data set value based on attribute and entities .The value of precision and recall which helps to detect the malware in various applications

3.6.1 Android Applications

Android application used to access the input values the application user wants to download.

3.6.1.1 Decompressor

It used to convert the text file into integer value.

3.6.1.2 Data Set

It uses matrix value for actual object and retrieved object and it separate the object into the corrected retrieved values after getting the values design the confused matrix .After getting data set matrix apply the precision and recall value. The data set value contain the actual object ,retrieved object and correct retrieved object .it include various applications like SMS,MMS,Webcam attribute or name of data set. Actual object that contain the value present in the applications, retrieved object include the positive predicted value and corrected value is all values which satisfies the actual and retrieved data parameters

Table 1. Data Set

Data Set Name	Objects		
	Actual Object	Retrieved Object	Corrected Retrieved Object
SMS	20	18	17
MMS	25	24	23

3.6.1.3 Precision

Precision is positive predictive value, is the fraction of retrieved instances that are relevant. Precision is the fraction of retrieved documents that are relevant to the find

Precision = | (relevant document) ∩ (retrieved document) | / (retrieved document)

It is the ratio of the number of relevant records retrieved to the total number of irrelevant and relevant records retrieved..[7] The precision value shows the relevant intersect to retrieved value to the total retrieved value which generate retrieved object.

Table 2. Precision

Precision	Objects		
	Actual Object	Corrected Retrieved Object	Value
SMS	17	18	0.944445
MMS	23	24	0.958334

3.6.1.4 Recall

It is the fraction of relevant instances that are retrieved. Recall in information retrieval is the fraction of the documents that are relevant to the query that are successfully retrieved.

Precision = | (relevant document) ∩ (retrieved document) | / (relevant document)

It is the ratio of the number of relevant records retrieved to the total number of relevant records in the database. It is usually expressed as a percentage. The recall value shows the relevant intersect to retrieved value to the total relevant value which generate actual object after that generation of corrected object which gives the value related to the application.

Table 3. Recall

Recall	Objects		
	Actual Object	Corrected Retrieved Object	Value
SMS	17	20	0.85
MMS	23	25	0.92

3.6.1.5 Matrix

The matrix represent the value of data set which all attributed are fulfilled ,it also contain the total no of attribute which contain all values must be satisfied represent as positive value represent one otherwise it is zero.

Table 4. Matrix

Matrix	Objects		
	Actual Object	SMS	MMS
	SMS	17	1
	MMS	1	18

3.6.1.6 Result

The result shows the values of both parameters i.e. precision and recall and average value of both the parameters.

Table 5. Result

Object	Precision	Recall
SMS	0.94445	0.85
MMS	0.95833	0.92

4. ARCHITECTURE

The design process of the system which includes the application program interfaces, input file which is to be transformed. The transformation contains the text to integer value conversion. Malware is detected after performing the algorithm over it, the generation of data set stored in database and that data set is apply towards the Naïve Bayes algorithm. It helps to find out the malware in the application like SMS, MMS, and Bluetooth. Result is calculated based on parameters as precision and recall ,the term precision and recall generate the value on the based on actual object retrieved and total object present .The generating the precision and recall value detect or predict the malware in the application .the final result is based on the average accuracy of precision . In figure 1 gives the detail about the entry of different apps in mobile device.

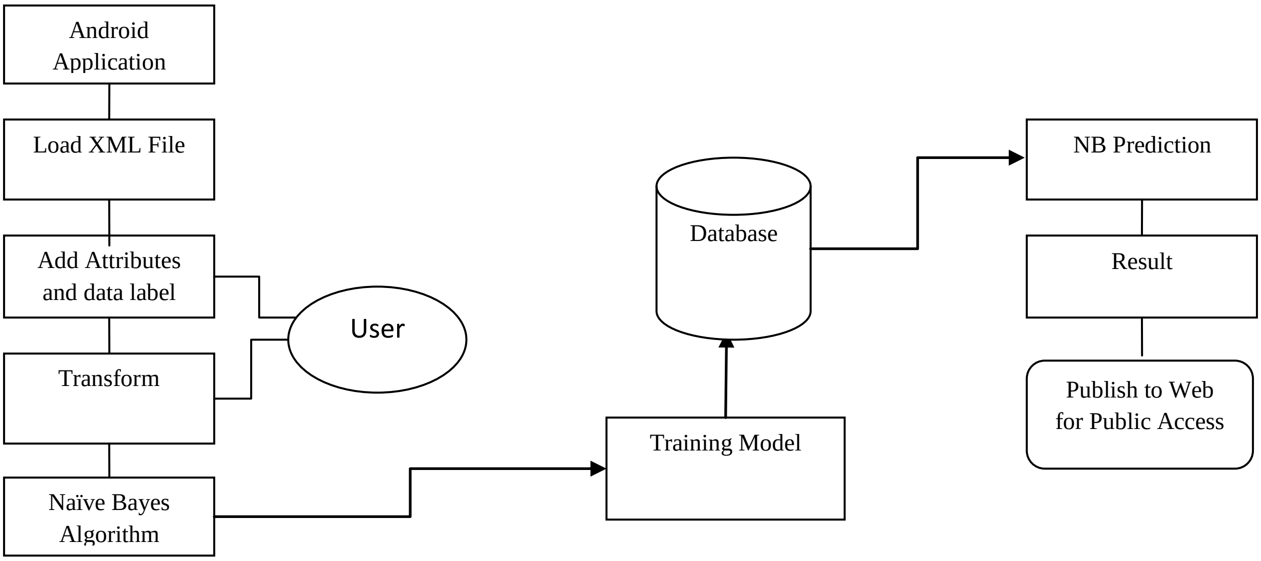

Fig 1: Malware Detection System Flow

5. CONCLUSION

Malware detection technique is based on data mining. This technique detects the malware, Trojans, virus from different sharing media like Bluetooth, MMS. The integrity defined from the security based smartphone. The smartphone security and providing integrity to mobile phone is big issue and it become overcome infected media, corruption of file, data lost problem related to the anti-virus program flow. Malware detection technique Naïve Bayes classifier uses the concept of data mining which helps to find the attributes in the application .Malware detection process helps to estimate the Trojans and virus report .It provide the security to the mobile user and application as well.

6. ACKNOWLEDGEMENT

We are thankful from JSPM'S BSIOTR and COMPUTER DEPARTMENT for their supports throughout the project stage. We would also like to thank the Head of Department and all faculties for their constant support and guidance .We also thank to our honorable Principal giving us motivation and facilities all the time.

7. REFERENCES

[1] Xinwen Zhang, Member, IEEE, Jean-Pierre Seifert, Member, IEEE, and Onur Aciic¸mez, Member, IEEE "Design and Implementation of Efficient Integrity Protection for Open Mobile Platforms," IEEE TRANSACTIONS ON MOBILE COMPUTING, VOL. 13, NO. 1, JANUARY 2014

[2] McAfee, "Mobile Security Report 2008," http://www.mcafee. com/us/research/mobile_security_report_2008.html, 2008

[3] D.D. Clark and D.R. Wilson, "A Comparison of Commercial and Military Computer Security Policies," Proc. IEEE Symp. Security and Privacy, 1987

[4] K.J. Biba, "Integrity Consideration for Secure Computer System, "Technical Report TR-3153, Mitre Corp., 1977.

[5] A.Bose and K. Shin, "Proactive Security for Mobile Messaging Networks," Proc. ACM Workshop Wireless Security, 2006

[6] G. Hu and D. Venugopal, "A Malware Signature Extraction and Detection Method Applied to Mobile Networks," Proc. IEEE 26th Int'l Performance, Computing, and Comm. Conf., 2007.

[7] Wikipedia ,"Precision and recall - Wikipedia, the free encyclopedia"

[8] www.google.com /"Naive Bayes Classifier"

[9] P. Loscocco and S. Smalley, "Integrating Flexible Support for Security Policies into the Linux Operating System," Proc. USENIX Ann. Technical Conf., pp. 29-42, June 2001