ISSN:0975-9646

# A Review on Risk Score Based App Classification Using Enriched Contextual Information of App Context

Lokhande Prajakta Padmakar[1], Prof. Shivaji R Lahane[2]

*Department of Computer Engg, University of Pune*

*GES's R. H. Sapat College of Engg, Nasik, India*

*Abstract*— **As the use of mobile devices is increasing rapidly day by day, huge number of mobile apps are coming into the market. Many of these apps are providing similar functionality and are coming from multiple unknown vendors. As a result, having a proper classification of these apps can turn out to be useful for various purposes like making it easy and time efficient for the user to select the required app, understanding the user preferences that can motivate the intelligent personalized services, etc. Having a effective classification for proper mobile app usage analysis, effective classification is required for which there is a need to have detailed information about the app. However this is nontrivial task as limited contextual information is available. As the information available about the apps is short and spares, classification of the apps can also be considered as coming into the category of classification of short and spares text. Various methods to classify the short and sparse text are present, which can be used for classification of the mobile apps. In this paper we have presented one such method in which we will be extracting the contextual information from sources like information from the labels (app name), information from the web search engine (snippets) and the contextual usage history of the app collected from the users usage record and also we will extract the permissions of the app requests before being installed. This is will give us effective and secure classification of the apps as most of these apps are coming from an unknown vendors and so they have the higher possibility of being malicious.**

*Keywords*— **Mobile app classification, Risk, Malware, Web knowledge, enriched contextual information, data mining.**

## I. INTRODUCTION

With the evolution in mobile technology, today the mobile phones are not limited to only calling and texting but they have moved a step ahead and have turned into an ubiquitous device. Now a days these devices provide different kind of facilities like camera, email facility, access to social network, video calling etc. As a result, varieties of mobile apps are coming into the market to provide these facilities. With easy to understand, tools made available it have become possible for anyone with little knowledge to design the mobile apps. Many of these apps provide us with similar kind of functionality, as a result having a classification of these apps will play an important role not only to the user in order to search the required app easily but also we can have the analysis of the user preferences which can help the intellectual services like app recommendation, user segmentation, target advertising etc.

For having a proper mobile app usage analysis the effective classification plays a very important role. Classification of these mobile apps is considered as a quite difficult task. This is because for having a proper or effective classification we need to have detailed information about the app. This is challenging task as very limited contextual information about the app is available. To be specific contextual information obtained from the apps name is very limited, as the words used for app name are very short and sparse. The objective of this paper is to design a system that will provide a effective classification of the mobile apps by using the enriched information about the apps.

To achieve this goal, we will be exploiting not only the web knowledge but also the real world contextual features about the apps along with their word labels. This will automatically improve the contextual information of the apps, resulting into improved performance of the classification. Here the web knowledge is extracted from the general search engine like google or from the app store, while the real world features will be extracted from the mobile usage record of the user.

Along with this information we also extract the permissions the app request before being installed. This is because it is found that malicious attack is possible through these apps using the permissions i.e. the permissions can access the user's private and sensitive information. Such a malicious app will not be detected by the current security mechanism provided by android, as the security mechanism provided is in standalone fashion i.e. the user is asked to make decision about accepting allowing the app to access the resources requested by reading the permissions. Which the users will easily ignore, as they mainly focus on the reviews and ratings of the app [9]. So we will be exploiting these permissions and calculate the risk score based on them. Providing an effective and secure classification of the mobile apps.

## II. LITERATURE SURVEY

The novel problem of automatically classifying the mobile apps can also be considered as a problem to classify the short and sparse text. Here we have considered some of the related existing research work with regards to the classification of these short and sparse texts and then the earlier work done to improve the security of the android. These works are presented as follows:

X. H. Phan et al [3] in their work have presented a general framework to process the short and sparse text documents on the web. They have focused mainly on data sparseness and synonyms/hyponyms by exploiting the hidden topics discovered from large scale external document collection i.e. universal data set. Here leveraging the hidden topics has improved the representation of the short and sparse text for classification. The semantic topics are the additional textual features integrated with the words to improve the classification.

M. Sahami and T.D. Heilman [4], in their work they have presented a similarity kernel function based approach to find the similarity between the short text. They have found that the traditional cosine similarity measures like for example cosine coefficient produce inadequate results like suppose we for the two short texts like “AI” and “Artificial intelligence” it will give the similarity as zero though both the terms are actually related to each other. According to the results of their work, they have proved that there approach can effectively measure the similarity between short text snippets which by exploiting the web search engine and provide greater context for the short texts.
Classifying the queries is an important task, as it is beneficial for a number of higher level tasks like web search and advertising matching. But search queries are usually short, thus carry insufficient information to provide accurate classification. A.Z. Broder et al [5] in their work have proposed a methodology for classifying these short queries using blind feedback technique. In which given a query its topic will be determined by the web searched results that will be returned for the query. The empirical evaluation performed by the authors proved that the methodology yields higher classification for the queries.

Discovering the users having similar interests can be used for various applications like recommendation, segmentation for market analysis, advertising etc. Compared to web based mining of user habits, mobile based mining is more powerful as it can capture rich contextual information while capturing the activity information, as powerful sensors are present in smart devices. H. Ma, H.Cao, O.Yang, E.Chen and J.Tian [6], in their work have proposed an approach which leverages search snippets to build vector space for both app usage and categories and classifies the app usage records using the cosine space distance.

Applications before being installed ask for the permissions to access some or the other kind of information from user’s device. Unknown to the user these apps may get access to some sensitive information present on the users device, which may be harmful to the user in case of the malicious apps using this information for its beneficial purpose. In order to make the user know what kind of data is being accessed by the apps they have installed, W.Enack et al [7], proposed a tracking system for real time privacy monitoring on smart phones. Which informs the user when the application may be trying to send sensitive data from the phone. But it does not defend against the security and monetary focused malware which send out spam or create premium SMS messages without accessing private information.

Applications downloaded from the smart phones request for much permission the user need to accept before they are being downloaded. A.P. Felt et al [8] studied these permissions and came to conclusion that most the apps asked for large number of permissions, which they don’t even require for their processing. This makes the apps more threatening as the attacker behind these apps may try to get access to the sensitive information of the user. Authors here have used the static analysis to determine the over privileged apps.

E. Chin et al [9] conducted a user study to gain insight into user perceptions of smart phone security and installation habits. Where it found that users don’t focus on the permissions during the app browsing and installation, they relay on the user rating and reviews while selecting the app.

TABLE I EXISTING WORK

| Sr No. | Authors | Work done |
|---|---|---|
| 1 | X. H. Phan et al | Proposed to leverage the hidden topics to improve the representation of the short and sparse text for classification. Here semantic topics are the additional textual features integrated with the words to improve the classification. |
| 2 | M. Sahami and T.D. Heilman | Proposed an approach for measuring the similarity between the short text snippets by exploiting the web search engine to provide greater context for the short texts. |
| 3 | A.Z. Broder et al | Proposed to build a robust query classification system which will identify thousands of query classes with reasonable accuracy. Blind feedback technique is used i.e. given a query its topic is determined by classifying the web search results retrieved by the query. Top related search results of the query are obtained from web search engine. |
| 4 | H. Ma, H.Cao ,O.Yang ,E. Chen and J.Tian | Proposed an approach which leverages search snippets to build vector space for both app usage and categories and classifies the app usage records using the cosine space distance |
| 5 | W.Enack et al | Proposed a tracking system for real time privacy monitoring on smart phones. here it informs the user when the application may be trying to send sensitive data off the phone |
| 6 | E. Chin et al | conducted a user study to gain insight into user perceptions of smart phone security and installation habit, concluded that users don’t focus on the permissions during the app browsing and installation, they relay on the user rating and reviews while selecting the app |
| 7 | A.P.Felt et al | Uses static analysis to determine if an android app is over privileged, i.e. if the is requesting for the permissions it never used. They found in the result that out of the 940 applications one third of the apps are over privileged |

## III. PROPOSED SYSTEM

In our proposed system to have an effective classification of the mobile apps, we will be exploiting and collecting information from various methods like web search engine, real world contextual data, contextual log information of users etc. From this data, we obtain the features for the apps appearing in these logs. Then with the help of machine learning model available, we will train the classifier to give us the appropriate classification of the app. To explain this in a more systematic way, consider given taxonomy T, and an app A and specified system parameter S, according to our approach, which will be extracted from the relevant web search and the contextual information about the app. To be more specific suppose we have app 'A' and then it will be classified into the S which consist of list of categories in order like $\{c_{1,}\ c_{2,}c_{3}....c_{s}\}$. Here for the effective feature selection an important task, is to train the machine learning model because the names of the apps are short and sparse as a result when a new app comes, whose partial or all the words present in the name are not present in the training data then the app will not be properly classified, so to overcome this we are extracting the features from different sources so that the relevance between this app and the categories can be obtained from these features. The system implemented will act as a standardization which can be used for various systems like app stores, target advertising, recommendation system, user segmentation etc. The system works according to the following phases.

### *A. Module 1: Feature extraction*

In this phase features of the app will be extracted from different sources. Our system considers both the explicit and implicit features of the app as explained below:

*1) Web based features:*

Here the both the explicit and implicit textual features from the web search engine will be extracted. To be specific, given the name of the app to the search engine the top most relevant results will be used to place the app in the defined category label. This is the explicit method in which the latent semantic meaning between the words is not considered. So as to have a more specific classification after having the explicit features of the app, we will consider the latent semantic meaning between the different words (implicit method).

*2) Contextual Features:*

Here the contextual information (features) from the real world for the app will be extracted. Here also both the explicit and implicit feedback will be considered.

In explicit feedback the feature-value pairs of the app are considered. To be specific for a app belonging to some category its feature-value pair like "day period" or "location", i.e. when that app is being used. Such type of context records are collected from the context logs of mobile users. A context profile for the same will be build; this will include the feature value pair and the frequency of its occurrence.

In implicit feedback the semantic meaning of the contextual feature-value pairs will be extracted from the pseudo feedback available. Like for example feature-value pairs like "period: evening", "holiday: yes" can be grouped together under the topic "relax" [1].

### *B. Module 2: Training the classifier:*

In this phase the machine learning model will be trained accordingly to combine the different features extracted from various sources for the apps. The classifier will be trained using the algorithms present to train the machine learning model. So that when the app will be given to the classifier it will effectively classify it into the category the app will belong.

### *C. Module 3: Contribution:*

To make this classification of the app more efficient and effective we will be extracting the meta information of the apps like the list of permissions they request to access from the user and also the usage statistics. Here using the rarity of the critical permission and the pairs of critical permissions used we will calculate the risk of the app in a simple user-friendly manner. The machine learning technique and heuristics, technique can be presented to generate the risk signals and risk score.

### *D. Module 4: Risk score based classification:*

Based on the overall category classification and the risk factor obtained for each app, the sysytem will give the final ranked classification of the apps. The apps will be ranked based on their risk factor in lower to higher order i.e. the apps with low risk factor will be placed first in the order and so on. The category wise classification will be helpful while analysing the user preferences which will improve the market analysis ultimately the target advertising, or user segmentation. On the other hand risk score will be help in identifying the apps having the low risk factor; this will improve the security concerns of android. This will help in improving the app ecosystem as with users paying more attention to the app will lower risk, the developers will follow the least privilege principle and will request only the necessary permissions for the app.

## IV. CONCLUSIONS

Effective classification of the mobile apps is important as everyday there are number of similar kind of apps coming in the market. Various classification techniques are present for classifying the short and sparse data, which can be adapted for classifying the mobile apps. But the result obtained from these techniques does not give us the effective classification of the apps[3][6]. As they take into consideration only single factor for classification i.e. web knowledge or contextual information .So we have proposed an approach to effectively classify the mobile apps in which we will extract the information from multiple sources in order to improve classification so as to provide more effective result. We have also considered leveraging the permissions requested by the apps which will then improve the ranking of the apps accordingly. Thus improving the security concerns of the malicious apps in easy to understand manner. This will not only help the user to select the proper app according to his requirements but also can be used for various other purposes like target advertising, user segmentation for market analysis etc.

## REFERENCES

[1] Hengshu Zhu, Enhong Chen, Hui Xiong, Huanhuan Cao, and Jilei Tian,``Mobile App Classification with Enriched Contextual Information ",IEEE Transactions on mobile computing (Volume:13 , Issue: 07 ),7 July 2014.

[2] Christopher S. Gates, Ninghui Li,Hao Peng, Bhaskar Sharma,Yuan Qi, Rahul Potharaju,Cristina Nita-Rotaru and Ian Molloy, ``Generating Summary Risk Scores For Mobile Applications",IEEE Transactions on dependable and secure computing (Volume :11, Issue: 03), May-June 2014

[3] X.-H. Phan et al., “A hidden topic-based framework toward building applications with short web documents,” IEEE Trans. Knowl.Data Eng., vol. 23, no. 7, pp. 961–976, Jul. 2010

[4] M. Sahami and T. D. Heilman, “A web-based kernel function for measuring the similarity of short text snippets,” in Proc. WWW, Edinburgh, U.K., 2006, pp. 377–386.

[5] Z. Broder et al., “Robust classification of rare queries using web knowledge,” in Proc. SIGIR, Amsterdam, Netherlands, 2007, pp. 231–238.

[6] H. Ma, H. Cao, Q. Yang, E. Chen, and J. Tian, “A habit mining approach for discovering similar mobile users,” in Proc. WWW, Lyon, France, 2012, pp. 231–240

[7] W. Enck, P. Gilbert, B. Chun, L.P. Cox, J. Jung, P. McDaniel, and A.N Sheth, “TaintDroid: An Information-Flow Tracking System for Realtime Privacy Monitoring on Smartphones,” Proc. Ninth USENIX Conf. Operating Systems Design and Implementation, article 1-6, 2010

[8] A.P. Felt, E. Chin, S. Hanna, D. Song, and D. Wagner, “Android Permissions Demystified,” Proc. 18th ACM Conf. Computer and Comm. Security, pp. 627-638, 2011.

[9] Chin, A.P. Felt, V. Sekar, and D. Wagner, “Measuring User Confidence in Smartphone Security and Privacy,” Proc. Eighth Symp. Usable Privacy and Security, (SOUPS ’12), article 1, 2012.