IARJSET
ISSN (Online) 2393-8021
ISSN (Print) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
ISO 3297:2007 Certified
Vol. 3, Issue 7, July 2016

Ranking Fraud Detection Using Opinion Mining for Mobile Apps

Tejaswini B. Gade[1], Prof. Nilesh G. Pardeshi[2]

PG Student, Computer Engineering Department, SRESCOE, Kopargaon, India[1]

Assistant Professor, Computer Engineering Department, SRESCOE, Kopargaon, India[2]

Abstract: Now days, Due to faster development in the mobile technology and mobile devices, the applications that is mobile apps are being very interesting and well-known concept in this field. As there is large number of mobile Apps, ranking fraud is the key challenge in front of the mobile App market. Ranking fraud is the term used for referring to fraudulent or suspicious activities which have intention of boosting up the Apps in the popularity list. In fact, App developers are using tricky means more and more frequently for increasing their Apps' sales or posting fake App ratings. Thus the need for preventing ranking fraud has been widely realized. This paper proposes a system for mobile apps in order to ranking fraud detection. The proposed system mines the leading sessions of mobile apps to precisely locate the ranking fraud. Additionally, system finds ranking, rating and review behaviors and investigation of three types of evidences, they are ranking based evidences, rating based evidences and review based evidences is done. Then, we propose an aggregation method based on optimization to combine all the evidences for fraud detection. Finally, the proposed system will be measured with App data collected from the App Store for a long time period.

Keywords: Mobile Apps, ranking fraud detection, evidence aggregation, historical ranking records, rating and review.

I. INTRODUCTION

Over the last few years the number of mobile Apps has been growing on a very large scale. At the end of April 2013 there is number of more than 1.6 million Applications at Apple's App store and Google Play. Different App stores launched their leader board on daily basis to inspire the development of mobile Apps which displays the chart rankings of most popular Apps. In fact for promoting mobile Apps, leader board of apps is the most important ways in the market. An app ranking at the top on the leader board ultimately leads to a large number of downloads and million dollars in revenue. This results in exploring of different ways by the App developers like organizing promotional drives to advertise their Apps in order to get top position in App leader boards.

The very recent trend followed in market by the corrupt App developers for bumping up of an App is to use deceptive means to intentionally boost their apps. Lastly, the chart rankings on a App store are also manipulated. This is usually implemented by using so-called "internet bots" or "human water armies" to raise the App downloads, ratings and reviews in a very little time. Venture Beat [1] is an article that reported, using ranking manipulation when an App was promoted, in Apple's top free leader board it could be push forward from number 1,800 to the upmost 25 and new users more than 50,000-100,000 could be acquired within a couple of days. In reality, such ranking fraud leads to great concerns to the industry of mobile App. For example, App developers who commit ranking fraud [2] in the App store, Apple has warned of cracking down on them. As per the observation the mobile apps does not always ranked high in the leader boards, in fact in some leading events only.

Collection of leading events of mobile Apps ultimately leads to different leading sessions. Thus, detecting ranking fraud of mob Apps happens in leading sessions and perhaps the process of detecting ranking fraud is done within the leading session of the mobile Apps. Especially, on the basis of historical ranking records of the mobile apps this paper proposes a simple and effective algorithm for the recognition of the leading sessions of each mobile App. This is one of the evidence collected from historical ranking records of apps against fraud. Moreover, there are two more types of fraud evidences proposed on the basis of Apps' rating and review history, which provides few anomaly patterns from Apps' historical rating and review records. Additionally, system propose an unsupervised evidence-aggregation method to combine these three types of evidences collected for the assessment of credibility of leading sessions from mobile Apps. At the end, the proposed system is evaluated with app data collected from various resources.

The rest of the paper contents are organized as follows: Section II the literature survey is presented over the related work. In section III, problem statement is mentioned, section IV presents the proposed system. Section V describes mathematical model and section VI shows the results of system. Finally, the section VII concludes the paper.

II. LITERATURE SURVEY

In this section, previous research papers related to the detection of ranking fraud for mobile Apps are studied. The research work of this study comprises of web ranking

IARJSET
ISSN (Online) 2393-8021
ISSN (Print) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
ISO 3297:2007 Certified
Vol. 3, Issue 7, July 2016

spam detection [3], [4], [5], online review spam detection [6], [7], [8] and mobile App recommendation [9], [10], [11], [12]. The web ranking spam refers to any purposeful actions which bring to selected webpages an inexcusable auspicious relevant importance [5]. Following is the work done on web ranking spam detection.

A. Ntoulaset al. [3] proposed various heuristic methods for content based spam detection. Different aspects of content based spam on the web are put forth by authors to find the heuristic methods.

N. Zhou et al. [5] presented the unsupervised web ranking spam detection. Spamicity was used by him for proposing an effective online link spam and methods for spam detection.

N. Jindal et al. [6] proposed methods for finding the users that generate spam reviews and identify their behaviors to model them in order to detect the spammers.

H. Zhu et al. [10] illustrates extraction of personal context-aware preferences device logs that is context logs for developing novel personalized context-aware recommender systems. Users download the application and installed it , however, is not only a indicator of whether user actually likes that application. Sometimes users only download and install the applications to try them out. So it becomes necessary to check the context logs of users to mine personal context-aware preferences of users.

Yong Geet al. [14] proposed the system for detection of taxi driving fraud which is committed by fraudulent taxi drivers to earn the money. They take unnecessary detours to passengers to commit the fraud by overcharging the passengers. So this kind of fraud are detected using GPS traces collected from number of taxi's and from these GPS traces various evidences are collected and finally these evidences are aggregated using dempster-shafer theory.

III.PROBLEM STATEMENT

The mobile industry is growing rapidly, subsequently the number of mobile apps coming in the market is also increasing. As there are many apps available in market users are confused while downloading the apps for their use. They check the daily app leader boards for selecting app. But few fraudulent app developers are using shady means for bumping up their apps on the leader board in order to get revenue. So detect such fraud apps we develop a system based on evidences i.e. Ranking fraud detection using opinion mining for mobile apps.

IV.PROPOSED SYSTEM

As there is increase in the number of mobile apps, fraudulent Apps must be detected; we have proposed a simple and effective algorithm for identifying the leading sessions of each App based on its historical ranking of records. With the analysis of ranking behaviors of Apps, we recognize that the fraudulent Apps often having different ranking patterns in their each leading session compared with normal Apps. Some fraud evidences are identified from Apps' historical ranking records resulting in development of three functions to detect likewise ranking based fraud evidences.

Moreover, two types of fraud evidences based on Apps' rating and review history are proposed. Fig. 1 depicts the framework of ranking fraud detection system for mobile Apps.

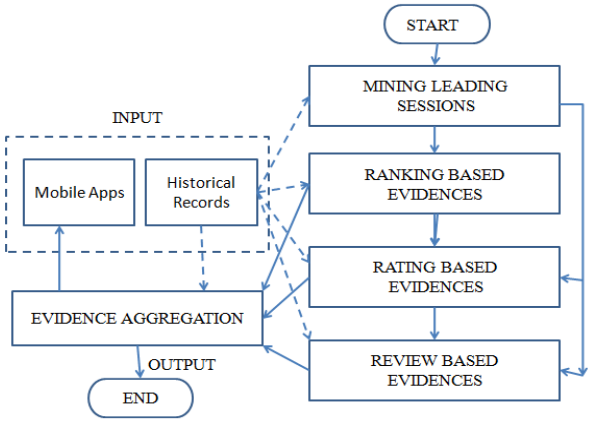

Fig. 1. Ranking Fraud Detection using Opinion Mining for Mobile Apps Overview

A. Rating based evidences

Rating to app is given by the user who downloaded it, specifically after the app is published in the market. Hence rating is one of the main evidence in ranking fraud of apps.In this module it performs preprocessing of ratings that is it removes ratings that are less than or equal to two and calculates rating score by by summing all the ratings class collected and decision is taken on the basis of rating which scores high amongst all.

B. Review based evidences

Reviews are familiar to all which provides the way for app user to write some textual comments regarding the personal experience of usage of that particular app.Therefore, manipulation of reviews is one way used by shady app developers to promote their app. Hence reviews are used to detect the ranking fraud in Mobile App industry. This module performs pre-processing of reviews and then performs sentiment analysis on pre-processed reviews.It will find out whether the comment is positive, negative or neutral. If word is positive then it will add plus one to score if word is negative it will minus one from score. Sometimes it is unable to find sentiment of some reviews, that time it makes the use of Naïve Bayes classifier.In this way it will find final score by analyzing sentiment of each review and determine whether app is fraud or not on the basis of review evidences.

C. Ranking based evidences

As per the observation the mobile apps does not always ranked high in the leaderboards, in fact in some leading events only. Further, App having adjacent leading events are merged to form leading sessions. Hence, the problem of identifying ranking fraud is to find out vulnerable leading sessions. There are two phases for mining leading sessions. Firstly, we need to discover the leading events

IARJSET
ISSN (Online) 2393-8021
ISSN (Print) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
ISO 3297:2007 Certified
Vol. 3, Issue 7, July 2016

from the historical ranking records of apps. Secondly, merging of adjacent leading events must be done for constructing leading sessions. Specially, Algorithm 1 demonstrates the pseudo code of finding leading sessions for a given App 'a' is.

Algorithm 1 Mining Leading Sessions
Input 1: a's historical ranking records R_a;
Input 2: the ranking threshold K*;
Input 3: the merging threshold ϕ;
Output: the set of a's leading sessions S_a;
Initialization: $S_a = \emptyset$;

1: $E_a = \emptyset$; e = ∅; s = ∅; $t^e_{start} = 0$;
2: for each i ∈[1, $|R_a|$] do
3: if $r^a_i \leq K^*$ and $t^e_{start} = = 0$ then
4: $t^e_{start} = t_i$;
5: else if $r^a_i > K^*$ and $t^e_{start} \neq 0$ then
6: //found one event;
7: $t^e_{end} = t_{i-1}$; e $<t^e_{start}, t^e_{end}>$;
8: if $|E_a| == \emptyset$then
9: $E_a \cup = e$; $t^s_{start} = t^s_{start}$; $t^s_{end} = t^s_{end}$;
10: else if $(t^s_{start} - t^s_{end}) < \phi$ then
11: //e* is the last leading event before e in E_a;
12: $E_a \cup = e$; $t^s_{end} = t^e_{end}$;
13: else then
14: //found one session;
15: s= $<t^s_{start}, t^s_{end}, E_a>$;
16: $S_a \cup = s$; $E_a = \emptyset$; s = ∅is a new session;
17: go to Step 7;
18: $t^e_{start} = 0$; e = ∅is a new leading event;
19: return S_a

In algorithm, e denotes leading events given in tuple as $<t^e_{start}, t^e_{end}>$ and sessions are denoted as tuple $<t^s_{start}, t^s_{end}, E^s>$ where E^s is set ofleading events in leading session. Step 2 to 7 are used extract individual leading events and step 8 to 16 are used to mine leading sessions. In this way we can easily find leading events and sessions of app.

D. Evidence Aggregation
After successful extraction of three types of evidences, the next step is combination of those evidences for ranking fraud detection. The final evidence score Ψ*(s) as a linear combination of all the existing evidences as equation given below.

$$\Psi^*(s) = \sum_{i=1}^{N_\Psi} w_i \times \Psi_i(s), \quad s.t. \sum_{i=1}^{N_\Psi} w_i = 1,$$

V. MATHEMATICAL MODEL

A. Set Theory
- Input Set : From the above definition, we get the input set(I), which contains a input text file.

I= Data about reviews, ratings and ranking of apps.
- Process Set : Consider a set of processes which are usedin this system.

P={P1,P2,P3,P4}
P1 = Finding review based evidences.
P2 = Finding rating based evidences.
P3 = Finding ranking based evidences.
P4 = Performing aggregation of all evidences.
- Intermediate Output Set : There are two output sets, The first is, intermediate output set is denoted by IO={IO1,IO2,IO3,IO4}

IO1 = Review based evidences generated.
IO2 = Rating based evidences determined.
IO3 = Ranking based evidences determined.
IO4 = Aggregation of evidences generated.
- Final Output Set:

O = Fraud detection of mobile app.

B. Venn Diagram
Venn diagram displays the mapping of the input, process andoutput of the system. It also represents the interaction between different processes along with input and output.

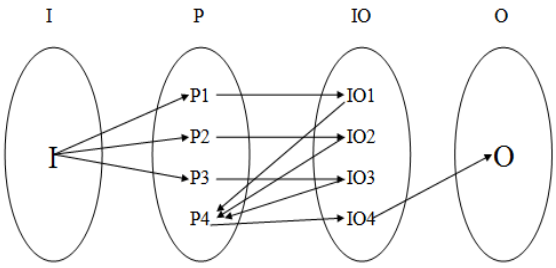

Fig.2. Venn diagram

C. Process State Diagram
Here, process p1, p2, p3 and process p4 are denoted by Q1, Q2, Q3 and Q4 respectively. Where Q5 is a final state that is Detection of whether app is fraud or not is shown in Fig.
3.

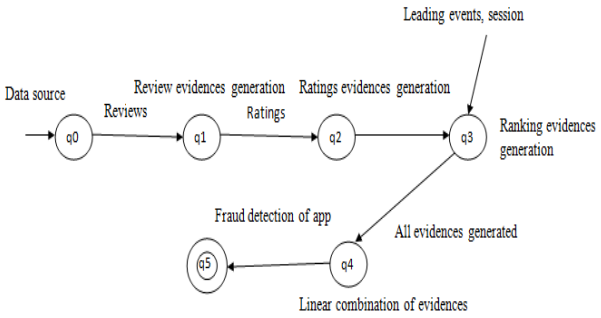

Fig. 3.Process State Diagram

VI. RESULT ANALYSIS

In this section performance evaluation is done to show the working efficiency of the proposed methodology. The experimental tests conducted were proving the effectiveness of the proposed methodology. In our work, varying number of apps is taken for analysis to predict the deceptive behavioral based apps. Using the proposed system each evidences are tested which shows its behavior

IARJSET
ISSN (Online) 2393-8021
ISSN (Print) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
ISO 3297:2007 Certified
Vol. 3, Issue 7, July 2016

in all types of evidences. Below given graphs shows behavior of apps in all types of evidences.

The performance evaluation on the basis of rating-based evidences is shown in the following Figure 4. It displays the count of positive and negative ratings of respective apps given in figure.

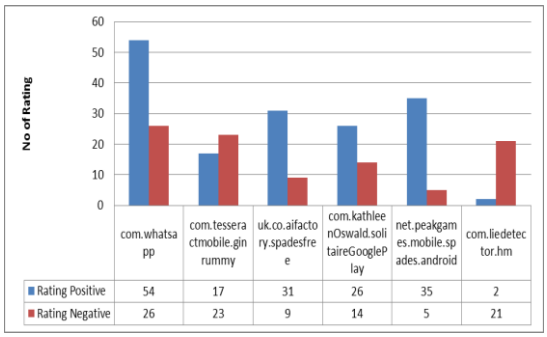

Fig. 4. Rating-Based Evidence Analysis

The performance evaluation on the basis of review-based evidences is shown in the following Figure 5. It displays the count of positive and negative reviews of respective apps.

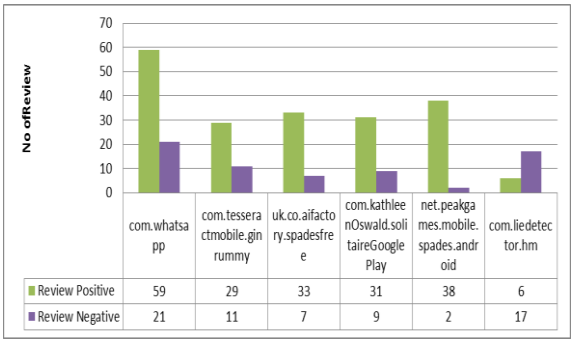

Fig. 5. Review-Based Evidence Analysis

The performance evaluation on the basis of ranking-based evidences is shown in the following Figure 6. It displays the average count of maintaining phase's i.e App Evidence-1 and also shows the respective session counts i.e App Evidence-2 of the apps in the figure 6. It isobserved that app having more number of sessions ultimately leads to low maintain phase. Hence the system considers that app as fraud with respect to ranking based evidences.

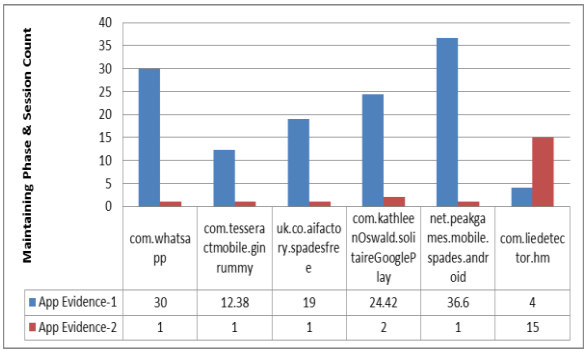

Fig. 6. Ranking-Based Evidence Analysis

From the above readings it is observed that, all evidences has its different output with respect to the selected apps.It is noticed that app with app_id com.tesseractmobile.ginrummy is having high negative rating count which is shown in figure 4, high positive review count which is shown in figure 5 and its maintaining phase and session count is also positive which is shown in figure 6. This shows that if we consider only rating based evidences then app is fraud and in review based evidences, ranking based evidences the app is geneuine. So the proposed system considers all three evidences while predicting any app as fraud or not. Hence when we aggregate all three evidences the system predicts that following app having id com.tesseractmobile.ginrummy is not fraud.

Now consider app with id com.liedetector.hm is having high negative rating count which is shown in figure 4, high negative review count which is shown in figure 5 and its maintaining phase and session count is also negative which is shown in figure 6. Hence when all three evidences are aggregated the system predicts that following app having id com.liedetector.hm is fraud.

In this way the proposed system does the prediction of app whether it is fraud or not on the basis of all the three evidences.

VII. CONCLUSION

This paper represents the novel approach for the development of a ranking fraud detection system for mobile apps. Firstly, identification of rating based evidences is done. Secondly, identification of review based evidences then by mining leading sessions ranking fraud evidences is collected. And finally system performs the aggregation of all three evidences to detect fraud apps. Experimental results showed the potency of the proposed approach. Our proposed system will definitely offer substantial benefits and provides an opportunity to prevent fraudulent apps being used in market.

ACKNOWLEDGMENT

I take his opportunity to express my hearty thanks to my guide Prof. N. G. Pardeshi for his guidance and sharing his findings for technical guidance and direction. Suggestions given by him were always helpful in this work to succeed. His leadership has been greatly valuable for me to work on this project and come with best out of it.

REFERENCES

[1] (2012). [Online]. Available: http://venturebeat.com/2012/07/03/apples-crackdown-on-app-ranking-manipulation/
[2] (2012). [Online]. Available: https://developer.apple.com/news/index.php?id=02062012a
[3] A. Ntoulas, M. Najork, M. Manasse, and D. Fetterly, "Detecting spam web pages through content analysis," in Proc. 15th Int. Conf. World Wide Web, 2006, pp. 83–92.
[4] N. Spirin and J. Han, "Survey on web spam detection: Principles and algorithms," SIGKDD Explor. Newslett., vol. 13, no. 2, pp. 50–64, May 2012.

IARJSET
ISSN (Online) 2393-8021
ISSN (Print) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
ISO 3297:2007 Certified
Vol. 3, Issue 7, July 2016

[5] B. Zhou, J. Pei, and Z. Tang, "A spamicity approach to web spam detection," in Proc. SIAM Int. Conf. Data Mining, 2008, pp. 277–288.
[6] E.-P. Lim, V.-A. Nguyen, N. Jindal, B. Liu, and H. W. Lauw, "Detecting product review spammers using rating behaviors," in Proc. 19thACMInt. Conf. Inform. Knowl.Manage., 2010, pp. 939–948.
[7] Z. Wu, J. Wu, J. Cao, and D. Tao, "HySAD: A semi-supervised hybrid shilling attack detector for trustworthy product recommendation," in Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2012, pp. 985–993.
[8] S. Xie, G. Wang, S. Lin, and P. S. Yu, "Review spam detection via temporal pattern discovery," in Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2012, pp. 823–831.
[9] K. Shi and K. Ali, "Getjar mobile application recommendations with very sparse datasets," in Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2012, pp. 204–212.
[10] B. Yan and G. Chen, "AppJoy: Personalized mobile application discovery," in Proc. 9th Int. Conf. Mobile Syst., Appl., Serv., 2011, pp. 113–126.
[11] H. Zhu, H. Cao, E. Chen, H. Xiong, and J. Tian, "Exploiting enriched contextual information for mobile app classification," in Proc. 21stACMInt. Conf. Inform. Knowl.Manage., 2012, pp. 1617–1621.
[12] H. Zhu, E. Chen, K. Yu, H. Cao, H. Xiong, and J. Tian, "Mining personal context-aware preferences for mobile users," in Proc. IEEE 12th Int. Conf. Data Mining, 2012, pp. 1212–1217.
[13] G. Shafer, A Mathematical Theory of Evidence. Princeton, NJ, USA: Princeton Univ. Press, 1976.
[14] Y. Ge, H. Xiong, C. Liu, and Z.-H. Zhou,"A taxi driving fraud detection system", in Proc. IEEE 11th Int. Conf. Data Mining, 2011.
[15] N. Jindal and B. Liu,"Opinion spam and analysis",in Proc. Int. Conf. Web Search Data Mining, 2008, pp. 219230.
[16] D. M. Blei, A. Y. Ng, and M. I. Jordan,"LatentDirichlet allocation" J. Mach. Learn.