International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2015): 78.96 | Impact Factor (2015): 6.391

Discovery of Historical Record to Detect Fraud App

Pritam Porate[1], M. S. Nimbarte[2]

[1]M. Tech Student, Department of Computer Science and Engineering, Bapurao Deshmukh College of Engineering, wardha, India

[2]Professor, Department of Computer Science and Engineering, Bapurao Deshmukh College of Engineering, wardha, India

Abstract: ***The confrontational growth of the mobile application market has made it a remarkable challenge for the users to find interesting applications in crowded App Stores. When users visit play store then they are able to see the various applications list. This list is built on the basis of promotion or advertisement. User doesn't have knowledge about the application (i.e. which applications are useful or useless). So user looks at the list and downloads the applications mostly from front page of App Store. But sometimes it happens that the downloaded application won't work or not useful. Furthermore, sent word dictionary is used to identify the exact reviews scores. The fake feedbacks by a same person for pushing up that app on the leaderboard are restricted. Two different constraints are considered for accepting the feedback given to an application. The first constraint is that an app can be rated only once from a user login. And the second is implemented with the aid of MAC address***

Keywords: Mobile Apps, Ranking Fraud Detection, Evidence Aggregation, Historical Ranking Records, Rating and Review, Recommendation app

1. Introduction

Ranking fraud in the mobile app market refers to fraudulent or deceptive activities which have a purpose of bumping up the apps in the popularity app developers to use shady means, such as inflating their apps ' sales or posting phony App ratings , to commit ranking fraud. While the importance of preventing ranking fraud has been widely recognized, there is limited understanding and research in this area. To this end, in this paper, we provide a holistic view of ranking fraud and propose a ranking fraud detection system for mobile apps. Specifically, we first propose to accurately locate the ranking fraud by mining the active periods, namely leading sessions, of mobile Apps. Such leading sessions can be leveraged for detecting the local anomaly instead of global anomaly of app rankings. Furthermore, we investigate three types of evidences, i.e., ranking based evidences, rating based evidences and review based evidences, by modeling apps ' ranking, rating and review behaviors through statistical hypotheses tests. In addition, we propose an optimization based aggregation method to integrate all the evidences for fraud detection.

The first is about web ranking spam detection. Particularly, the web ranking spam refers to any deliberate Actions which bring to selected webpages an unjustifiable Favorable relevance or importance. For example, Ntoulaset al. [2] have studied various aspects of content-based spam on the web and presented a number of heuristic methods for detecting content based spam. Zhou et al. [6] have studied the problem of unsupervised web ranking spam detection. Specifically, they proposed an efficient online link spam and term spam detection methods using spamicity. Recently, Spirin and Han [8] have reported a survey on web spam detection, which comprehensively introduces the principles and algorithms in the literature. Actually, the work of web ranking spam detection is mainly based on the analysis of ranking principles of search engines, like Page Rank and query term frequency. This is different from ranking fraud detection for mobile Apps. The second category is concentrated on detecting online review spam. For example, Lim et al. [9] have identified several indicative behaviors of review spammers and model these behaviors to detect the spammers. [5]A Semantic Association Page Rank Algorithm for Web Search Engines Manuel Rojas This paper propose a relation-based page rank formula to be used as a Semantic Web search engine. Wu et al. [10] have studied the problem of detecting hybrid shilling attacks on rating data. The proposed approach is based on the semi supervised learning and can be used for reliable product recommendation. Xie et al. [11] have studied the problem of singleton review spam detection. Specifically, they solved this problem by detecting the co-anomaly patterns in multiple review based time series. [14]Although some of above approaches can be used for anomaly detection from historical rating and review records, they are not able to extract fraud evidences for a given time period (i.e., leading session). Finally, the third category includes the studies on mobile App recommendation. For example, Yan and Chen [12] developed a mobile App recommender system, named Appjoy, which is based on user's App usage records to build a preference matrix instead of using explicit user ratings. Also, to solve the sparsity problem of App usage records, Shi and Ali [13] studied several recommendation models and proposed a content based collaborative filtering model, named Eigenapp, for recommending Apps in their website Getjar. In addition, some researchers studied the problem of exploiting enriched contextual information for mobile App recommendation. For example, Zhu et al. [14] proposed a uniform framework for personalized context-aware recommendation, which can integrate both context independency and dependency assumptions. However, to the best of our knowledge, none of previous works has studied the problem of ranking fraud detection for mobile Apps. In this paper, built up a positioning extortion identification framework for versatile applications that positioning misrepresentation happened in driving sessions for each application from its verifiable positioning records.[15] Ntoulaset al. have studied various aspects of content-based spam on the web and presented a number of heuristic methods for detecting content based spam[7].

2. Literature Survey

A. D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation" [1]:

D. M. Blei, A. Y. Ng, and M. I. Jordan, introduces a unique model called as Dirichlet allocation (LDA) a generative probabilistic model for collections of discrete data such as text amount. Basically it is a three level hierarchical Bayesian model in which each element of a group is demonstrated as a finite mixture over a fundamental set of topics. Each topic is demonstrated as an infinite mixture over fundamental set of topic probabilities. With the reference of text modelling, the topic probabilities provide an open representation of a document. An efficient approximation inference technique is presented based on various methods and an EM algorithm for empirical Bayes parameter estimation is also presented. The results are reported in document modelling, text classification and collaborative filtering, which compares to a collection of unigrams and probabilistic LSI model

B. Y. Ge, H. Xiong, C. Liu, and Z.-H. Zhou, "A Taxi Driving Fraud Detection Systemices in City Taxis" [2]:

Y. Ge, H. Xiong, C. Liu, and Z.-H. Zhou , illustrated that growth in the field of GPS tracking technology have allowed the users to install GPS tracking devices in taxies to gather huge amount of GPS traces under some time period. These traces by GPS offered an unparalleled opportunity to uncover taxi driving fraud traces and then fraud detection system is proposed which is able to identify taxi driving fraud. First, two sort of function are uncovered here i.e. travel route evidence and driving distance evidence. Even a third function is developed to combine the previous functions based on Dempster-Shafer theory. First identification of interesting locations is done from tremendous amount of taxi GPS logs and then parameter free method isproposed to extract the travel route evidences. Secondly, concept of route mark is +++developed to illustrate the driving path between locations and based on those mark, specific model is characterized for the distribution of driving distance and discover the driving distance evidences. And finally, taxi driving fraud detection system with
a large scale real world taxi GPS logs.

C. T. L. Griffiths and M. Steyvers, "Rank Aggregation Via Nuclear Norm Minimization" [3]:

T. L. Griffiths and M. Steyvers, introduces the process of rank aggregation which is interweave with the structure of skewsymmetric matrices. Recent development in the principles of matrix completion matrices is been applied and this idea gives rise to a new method for ranking a set of items. The core of this idea deals with the raking aggregation method which intimately describes a partially filled skew-symmetric matrix. The algorithm for matrix completion is raised to hold skew-symmetric data and use that to extract ranks for each item. This algorithm applies same strategy for both pairwise comparisons as well as for rating data. It becomes robust to noise and incomplete data as it is based on matrix completion.

D. A. Klementiev, D. Roth, And K. Small, "An Unsupervised Learning Algorithm for Rank Aggregation" [6]:

A. Klementiev, D. Roth, and K. Small, describes the field of information retrieval, data mining, and natural language, many applications needs a ranking of instances which is not present in classification. Furthermore, a rank aggregation is a result of aggregating the results of the established ranking models into formalism and then result represents a novel unsupervised learning algorithm (ULARA) which gives a linear combination of individual ranking functions. These functions were developed based on the axiom of rewarding ordering agreement between the rankers.

E. A. Klementiev, D. Roth, And K. Small, "Unsupervised Rank Aggregation with Distance-Based Models" [8]:

A. Klementiev, D. Roth, and K. Small, produces a model which has to integrate the set of rankings often deals with aggregating and it only comes up when a certain ranked data is developed. Even though the various heuristic and supervised learning approaches to rank aggregation, a requirement of domain knowledge and supervised ranked data exists. Therefore, to solve this issue, a framework is proposed for learning aggregate rankings without supervision. This framework is instantiated for the cases of permutations and combinations of top-k lists.

3. Proposed Work

Proposed System

In this paper, we propose to develop a ranking fraud detection system for mobile Apps. Along this line, we identify several important challenges. First, ranking fraud does not always happen in the whole life cycle of an App, so we need to detect the time when fraud happens. Such challenge can be regarded as detecting the local anomaly instead of global anomaly of mobile Apps. Second, due to the huge number of mobile Apps, it is difficult to manually label ranking fraud for each App, so it is important to have a scalable way to automatically detect ranking fraud without using any benchmark information. Finally, due to the dynamic nature of chart rankings, it is not easy to identify and confirm the evidences linked to ranking fraud, which motivates us to discover some implicit fraud patterns of mobile Apps as evidences.

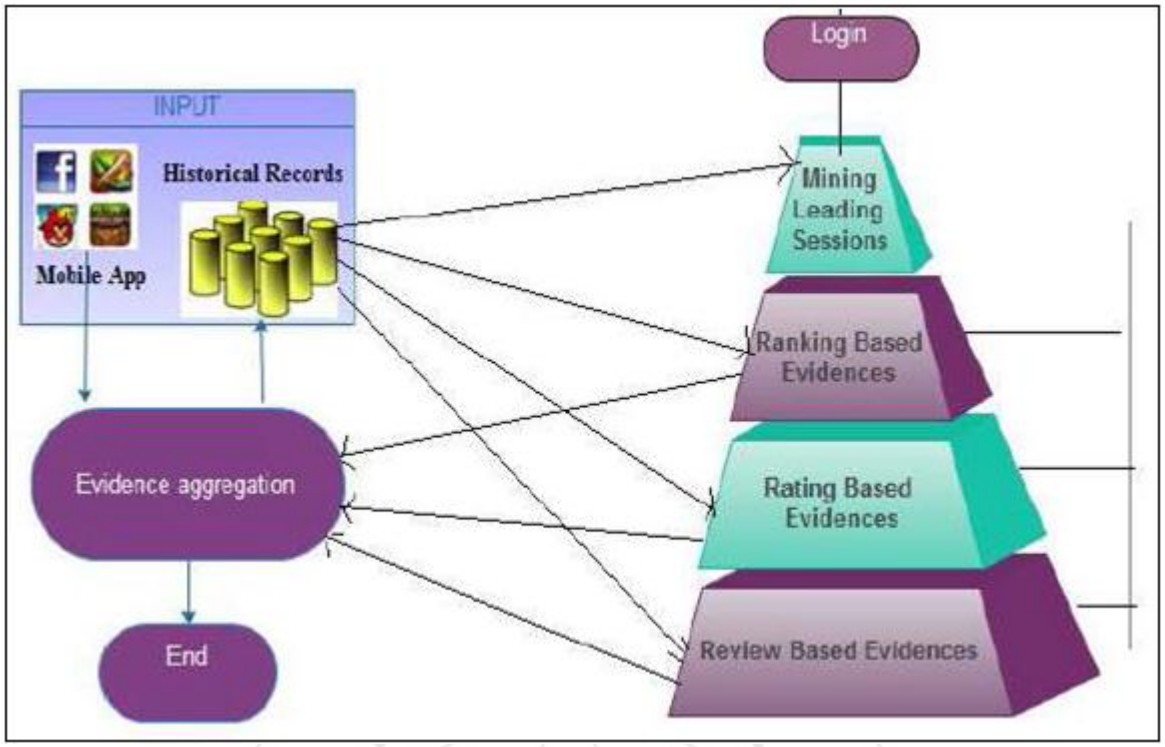

Figure 1: System Model

a. ***Ranking Based Evidences***

By analysing the Apps' historical ranking records, we observe that Apps' ranking behaviours in a leading event always satisfy a particular ranking pattern, which include the three dissimilar ranking phases, namely, rising phase, maintaining phase and recession phase. Particularly, in every one leading event, an App's ranking first increases to a peak position in the leader board (i.e., rising phase), at that time keeps such peak position for a period (i.e., maintaining phase), and lastly lessen till the end of the event (i.e., recession phase). Fig. 3 shows an example of different ranking phases of a leading event. In addition, such a ranking pattern shows a fundamental understanding of leading event. In the following, we formally define the three ranking phases of a leading event.

b. Rating Based Evidences

For ranking fraud detection are uses the ranking based evidences. However, sometimes, it is not sufficient to only use ranking based evidences. For instance, some Apps developed by the famous developers, such as Game loft, may have some leading events with large values of u1 due to the developers' credibility and the "word-of-mouth" advertising effect. Additionally, some of the legal marketing services, such as "limited-time discount", may also result in significant ranking based evidences. To solve that problem, we additionally study how to extract fraud evidences from Apps' historical rating records. Specifically, after an App has been published, it can be rated by any user who downloaded it. Indeed, user rating is one of the most valuable features of App advertisement. An App which has higher rating may attract more users to download and also can gives ranked higher in the leader board. Thus, rating manipulation is also a valuable perspective of ranking fraud. Innocently, if an App has ranking fraud in a leading session s, the ratings during the time period of s may have inconsistency patterns merged with its historical ratings, which can be used for constructing rating based evidences.

c. Review Based Evidences

In addition ratings, most of the App stores also permit users to write some textual comments as App reviews. Such reviews can indicates the individual perceptions and usage experiences of existing users for particular mobile Apps. Indeed, review manipulation is one of the most valuable perspective of App ranking fraud. Specifically, before downloading or purchasing a new mobile App, users usually first read its historical reviews to ease their decision making, and a mobile App contains more encouraging reviews may captivate more users to download. Therefore, imposters often post fake reviews in the leading sessions of a particular App in order to increases the App downloads, and thus propel the App's ranking position in the leaderboard. For all that previous works on review spam detection have been reported in recent years [4], [9], the issue of detecting the local inconsistency of reviews in the leading sessions and capturing them as evidences for ranking fraud detection are still under explored. For this purpose, here we propose two fraud evidences for detecting ranking fraud based on Apps' review behaviours in leading sessions.

d. Evidence Aggregation

After extracting all three types of fraud evidences, then the next challenge is how to combine them for ranking fraud detection. In addition, there are many methods of ranking and evidence aggregation in the literature, such as permutation based models [7], [8], score based models [11], and Dempster Shafer rules [10], [12]. However, some of these methods focus on learning a global ranking for all applicants. This way is not proper for detecting ranking fraud for new Apps. Distinct methods are based on supervised learning techniques, which rely on the labelled training data and are hard to be exploited. Rather, we

suggest an unsupervised approach based on fraud similarity to combine these evidences.

Detecting ranking fraud of mobile Apps is actually to detect ranking fraud within leading sessions of mobile Apps. Specifically, we first propose a simple yet effective algorithm to identify the leading sessions of each App based on its historical ranking records. Then, with the analysis of Apps' ranking behaviors, we find that the fraudulent Apps often have different ranking patterns in each leading session compared with normal Apps. Thus, we characterize some fraud evidences from Apps' historical ranking records, and develop three functions to extract such ranking based fraud evidences.

4. Conclusions

We conclude that, to develop a ranking fraud detection system for mobile Apps. we first discover that ranking fraud occur in leading sessions and provided a method for mining leading sessions for each App from its historical ranking records. In that case, we identified ranking based evidences, rating based evidences and review based evidences for detecting ranking fraud. Furthermore, we proposed an optimization based aggregation method to integrate all the evidences for evaluating the reliability of leading sessions from mobile Apps. That all the evidences can be modeled by statistical hypothesis tests for the unique perspective of this approach, thus it is easy to be extended with other evidences from domain knowledge to detect ranking fraud. Ultimately, we validate the proposed system with extensive experiments on real world App data collected from the Google play store. Experimental results showed the effectiveness of the proposed approach.

References

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," J. Mach. Learn. Res., pp. 993–1022, 2003.
[2] 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, pp. 181– 190.
[3] D. F.Gleich and L.-h. Lim, "Rank aggregation via nuclear norm minimization," in Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2011,pp. 60–68.
[4] T. L. Griffiths and M. Steyvers, "Finding scientific topics," Proc. Nat. Acad. Sci. USA, vol. 101, pp. 5228–5235, 2004.
[5] G. Heinrich,Parameter estimation for text analysis, " Univ. Leipzig, Leipzig, Germany, Tech. Rep., http://faculty.cs.byu.edu/~ringger/ CS601R/papers/Heinrich-GibbsLDA.pdf, 2008.
[6] B. Zhou, J. Pei, and Z. Tang. A spamicity approach to web spam detection. In *Proceedings of the 2008 SIAM International Conference on Data Mining*, SDM'08, pages 277–288, 2008.
[7] A. Ntoulas, M. Najork, M. Manasse, and D. Fetterly. Detecting spam web pages through content analysis. In *Proceedings of the 15th international conference on World Wide Web*, WWW '06, pages 83–92, 2006.
[8] N. Spirin and J. Han. Survey on web spam detection: principles and algorithms. *SIGKDD Explor. Newsl.*, 13(2):50–64, May 2012.
[9] E.-P. Lim, V.-A. Nguyen, N. Jindal, B. Liu, and H. W. Lauw. Detecting product review spammers using rating behaviors. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, CIKM '10, pages 939–948, 2010.
[10] Z.Wu, J.Wu, J. Cao, and D. Tao. Hysad: a semi-supervised hybrid shilling attack detector for trustworthy product recommendation. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '12, pages 985–993, 2012
[11] S. Xie, G. Wang, S. Lin, and P. S. Yu. Review spam detection via temporal pattern discovery. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '12, pages 823–831, 2012.
[12] B. Yan and G. Chen. Appjoy: personalized mobile application discovery. In *Proceedings of the 9th international conference on Mobile systems, applications, and services*, MobiSys 11, pages 113–126, 2011.
[13] K. Shi and K. Ali. Getjar mobile application recommendations with very sparse datasets. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '12, pages 204–212, 2012.
[14] H. Zhu, E. Chen, K. Yu, H. Cao, H. Xiong, and J. Tian. Mining personal context-aware preferences for mobile users. In *Data Mining (ICDM), 2012 IEEE 12th International Conference on*, pages 1212–1217, 2012.
[15] Hengshu Zhu, Hui XiongDiscovery of Ranking Fraud for Mobile Apps. IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING,2013.