MOBILE APP FRAUD DETECTION SYSTEM

D. Saritha, Asst. Professor, Dept.of Master of Computer Applications, Narayana Engineering College(Autonomous), Gudur.SPSR Nellore, AP, India.
P. Joshna, PG scholar, Dept.of Master of Computer Applications, Narayana Engineering College(Autonomous), Gudur.SPSR Nellore, AP, India.

Abstract- Mobile App ranking fraud refers to fraudulent or deceptive activities that have the purpose of boosting Apps' popularity rankings. App developers are increasingly resorting to shady tactics, such as inflating their Apps' sales or posting phoney App ratings, to manipulate rankings. Despite widespread recognition of the importance of preventing ranking fraud, research and understanding in this area are limited. A ranking fraud detection system for mobile apps is proposed in this paper. We propose, in particular, to locate ranking fraud accurately by mining the active periods, i.e., the leading sessions, of mobile applications. The local anomaly of App rankings can be detected using such leading sessions. A statistical hypotheses test is used to model the Apps' ranking, rating, and review behaviours in order to examine three types of evidence. Additionally, we propose a method of aggregation based on optimization for fraud detection. Our final evaluation involves real-world App Store data. This system has been validated in experiments, which also demonstrate the scalability of detection algorithms and a degree of regularity in ranking fraud activities.

1.INTRODUCTION

On day by day basis, an app leaderboard can be up to date by way of app shop which shows chart rankings of most famous apps, additionally it is an inspirational component to make stimulated the improvement of cell apps. In fact, for merchandising cell Apps, chief board of apps is the most essential way of upgradient in the market. An app ought to be ranked greater rely upon how its chart of improvement increase and step by step it can make wide variety of downloads and finally excessive income in dollar. There had been unique methods to promote Apps promotional power in order to get pinnacle function in App leaderboards the criminal one is white hat foundation to promote their App to get well-known and alternately extra quantity of downloads. But there are additionally some unlawful methods say black hat groundwork for bumping up the App via the use of some misleading skill used by means of corrupt App builders to get well-known in some quick time period. This method commonly carried out via the use of so-called “internet bots” or “human water armies” to increase the App downloads, rankings and opinions in a very little time. Some are crucial factors that is to avert fraud, confirmed as given two constraints. The first constraint is that an app can be rated only as soon as from a consumer login and the 2d is applied with the useful resource of IP tackle that limits the variety of consumer login logged per day. Finally, the proposed gadget will be evaluated with realworld App statistics which is to be accumulated from the App Store for a long-time duration referred to as historic records. In the present system, from the accrued historical records, the main tournament and main session of an app is identified. There are two principal steps for mining main sessions. First, we want to find out main occasions from the App’s historic rating records. Second, we want to merge adjoining main activities for establishing main sessions. Careful commentary indicates that the cell Apps are now not continually at pinnacle most function in chief board. But solely in some time length referred to as main tournament which is shape one of a kind main periods ability rating fraud mainly manifest in this main session. Then from the consumer judgmental feedback, three specific sorts of evidences are gathered specifically rating primarily based evidence, ranking primarily based proof and assessment based totally evidence. As our challenge based totally on evidences gathered from app data; the one of the primarily judgment by using human beings is ranking primarily based evidences which can be used to price the app whilst downloading it or we can fee it after seeing its performance. It is most necessary proof to choose the app. But as mentioned above there are some methods with assist of which the ranking can get will increase via doing fraud. So, some other judged proof based totally approach is assessment based

totally evidence; discovering to make the genuine specification of app whether or not it is true or horrific app to download. In Review Based Evidences, without ratings, most of the App shops additionally enable customers to write some textual remarks as App reviews. So, human beings may also certain shot about downloading that unique app by means of studying feedback detailed in overview area and additionally supply their opinion about that app. Due to the big variety of apps, it is difficult to search rating fraud for each apps; so, it is necessary to have a scalable way to routinely become aware of rating fraud except the use of any benchmark Research information. So right here are come the notion of algorithm used in our project. Especially, this paper proposes a easy and tremendous algorithm to understand the main classes of every cellular App primarily based on its historic rating records. Here we are proposing some statistical check the place the records supply the specific demonstration of things to do executed by way of the app to rank itself. If rank is maintained over length and layout falls and so a whole lot fluctuations can be viewed then these apps have to desire strong verification to location it in remaining role or make it out of the play store. Also, ranking and evaluation history, which offers some anomaly patterns from apps historic ranking and opinions records. Also, we are making right here semantic evaluation check over these accrued data; for example, right here in this check we use evaluation to discover positive, terrible and impartial consequences on these feedback and judging app up to the mark

2.LITERATURE SURVEY

1) A flexible generative model for preference aggregation

AUTHORS: M. N. Volkovs and R. S. Zemel

Many areas of study, such as information retrieval, collaborative filtering, and social choice face the preference aggregation problem, in which multiple preferences over objects must be combined into a consensus ranking. Preferences over items can be expressed in a variety of forms, which makes the aggregation problem difficult. In this work we formulate a flexible probabilistic model over pairwise comparisons that can accommodate all these forms. Inference in the model is very fast, making it applicable to problems with hundreds of thousands of preferences. Experiments on benchmark datasets demonstrate superior performance to existing methods .

2) Getjar mobile application recommendations with very sparse datasets

AUTHORS: K. Shi and K. Ali

The Netflix competition of 2006 [2] has spurred significant activity in the commendations field, particularly in approaches using latent factor models [3,5,8,12] However, the near ubiquity of the Netflix and the similar MovieLens datasets1 may be narrowing the generality of lessons learned in this field. At GetJar, our goal is to make appealing recommendations of mobile applications (apps). For app usage, we observe a distribution that has higher kurtosis (heavier head and longer tail) than that for the aforementioned movie datasets. This happens primarily because of the large disparity in resources available to app developers and the low cost of app publication relative to movies.

In this paper we compare a latent factor (PureSVD) and a memory-based model with our novel PCA-based model, which we call Eigenapp. We use both accuracy and variety as evaluation metrics. PureSVD did not perform well due to its reliance on explicit feedback such as ratings, which we do not have. Memory-based approaches that perform vector operations in the original high dimensional space over-predict popular apps because they fail to capture the neighborhood of less popular apps. They have high accuracy due to the concentration of mass in the head, but did poorly in terms of variety of apps exposed. Eigenapp, which exploits neighborhood information in low dimensional spaces, did well both on precision and variety, underscoring the importance of dimensionality reduction to form quality neighborhoods in high kurtosis distributions.

3) Detecting spam web pages through content analysis

AUTHORS: A. Ntoulas, M. Najork, M. Manasse, and D. Fetterly

In this paper, we continue our investigations of "web spam": the injection of artificially-created pages into the web in order to influence the results from search engines, to drive traffic to certain pages for fun or profit. This paper considers some previously-undescribed techniques for

automatically detecting spam pages, examines the effectiveness of these techniques in isolation and when aggregated using classification algorithms. When combined, our heuristics correctly identify 2,037 (86.2%) of the 2,364 spam pages (13.8%) in our judged collection of 17,168 pages, while misidentifying 526 spam and non-spam pages (3.1%).

4) Spotting opinion spammers using behavioral footprints

AUTHORS: A. Mukherjee, A. Kumar, B. Liu, J. Wang, M. Hsu, M. Castellanos, and R. Ghosh

Opinionated social media such as product reviews are now widely used by individuals and organizations for their decision making. However, due to the reason of profit or fame, people try to game the system by opinion spamming (e.g., writing fake reviews) to promote or to demote some target products. In recent years, fake review detection has attracted significant attention from both the business and research communities. However, due to the difficulty of human labeling needed for supervised learning and evaluation, the problem remains to be highly challenging. This work proposes a novel angle to the problem by modeling spamicity as latent. An unsupervised model, called Author Spamicity Model (ASM), is proposed. It works in the Bayesian setting, which facilitates modeling spamicity of authors as latent and allows us to exploit various observed behavioral footprints of reviewers. The intuition is that opinion spammers have different behavioral distributions than non-spammers. This creates a distributional divergence between the latent population distributions of two clusters: spammers and non-spammers. Model inference results in learning the population distributions of the two clusters. Several extensions of ASM are also considered leveraging from different priors. Experiments on a real-life Amazon review dataset demonstrate the effectiveness of the proposed models which significantly outperform the state-of-the-art competitors.

3. PROPOSED WORK

- 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.
- We further propose two types of fraud evidences based on Apps' rating and review history, which reflect some anomaly patterns from Apps' historical rating and review records.
- In Ranking Based Evidences, by analyzing the Apps' historical ranking records, we observe that Apps' ranking behaviors in a leading event always satisfy a specific ranking pattern, which consists of three different ranking phases, namely, rising phase, maintaining phase and recession phase.
- In Rating Based Evidences, 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 important features of App advertisement. An App which has higher rating may attract more users to download and can also be ranked higher in the leaderboard. Thus, rating manipulation is also an important perspective of ranking fraud.
- In Review Based Evidences, besides ratings, most of the App stores also allow users to write some textual comments as App reviews. Such reviews can reflect the personal perceptions and usage experiences of existing users for particular mobile Apps. Indeed, review manipulation is one of the most important perspective of App ranking fraud

3.1 IMPLEMENTATION

Ranking fraud detection have proofs for detection of fraud; such proofs are evidences referred to as as rating based, evaluate and ranking primarily based evidences. These evidences are used to mine main sessions. Ranking based, evaluate and ranking based totally evidences are utilized

step via step. The precise rating sample is fulfilled by using app ranking conduct in rating primarily based evidences. In ranking based totally evidences, ranking sample that is rankings given by using consumer is used for ranking fraud detection. Rating is given by means of consumer at the time of downloading the App or after judging that App through the usage of it after some time. If the rankings are excessive for that App in greater extent or extra App customers supply nearly excessive rating then that App is attracted through greater cellular users. In this there are greater probabilities of make fraud through App developer with the aid of earn rankings carried out in main sessions. In assessment based totally evidences, opinions are feedback given via cellular app customers given after judging that app after downloading it. But right here earlier than cell app customers downloading that App they will sincerely goes via these feedback to get others view to clear their way to down load that App or not. As the quantity of cell Apps will increase day by means of day, fraudulent Apps have to be detected. So we have proposed a easy and superb algorithm for figuring out the main periods of every App primarily based on its historic rating of records. We apprehend that the fraudulent Apps frequently have unique rating patterns in every main session, in contrast with ordinary Apps in accordance to rating behaviors of Apps. Some fraud evidences are recognized from Apps historic rating records, which outcomes in improvement of three features to become aware of likewise rating primarily based fraud evidence. So moreover, right here two kinds of fraud evidences primarily based on Apps ranking and assessment records are proposed.

4. IDENTIFYING EVIDENCES FOR RANKING FRAUD DETECTION:

1. Identifying Leading Sessions: Leading sessions are the base for detecting fraud in mobile App as ranking fraud usually happens in leading sessions. And hence detecting ranking fraud is actually detecting ranking fraud within leading session of mobile Apps which we mine from mobile Apps historical ranking records. There are two main steps for mining primary sessions. First, we need to determine leading measures from the App's previous ranking records. Second, we need to collaborate neighboring leading events for developing leading sessions. Specifically, we first propose a simple yet effective algorithm to identify the leading events of each App based on its historical ranking records. Then, we merge adjacent leading events for constructing leading sessions. As per the observation the mobile apps do not always ranked high in the leader boards, in fact in some leading events only. With the analysis of Apps' ranking behaviors, the fraudulent Apps often have different ranking patterns in each leading session compared with normal Apps. Therefore, the problem of identifying ranking fraud is to find out vulnerable leading sessions.

2. Ranking based evidences: A leading session is composed of several leading events. Therefore, we should first analyze the basic characteristics of leading events for extracting fraud evidences. By analyzing the Apps' historical ranking accounts, Apps' ranking behaviors in a leading incident always assure a specific ranking pattern, which consists of three different ranking segments, expanding phase, maintaining phase and collapse phase. Mainly, in each leading event, an App's ranking first improve to a peak or extent position in the leaderboard (i.e., rising phase), then maintain such peak position for a phase (i.e., maintaining phase), and at last declines till the end of the event (i.e., recession phase). Definitely, such a ranking pattern confirms a significant consideration of leading event. In next section, we formally describe the three ranking phases of a leading event.

3. Rating based evidences:

The ranking based evidences are first step towards ranking fraud recognition. However, sometimes, it is not satisfactory to only use ranking based evidences. Take an example, some Apps formed by the legendary developers, such as Gameloft, may have some leading events due to the developers' trustworthiness and the "word-ofmouth" advertising effect. Moreover, some of the permissible marketing services, such as "limited-time discount", may also consequence in significant ranking based evidences. To solve this matter, we also study how to extort fraud evidences from Apps' historical previous rating records. Indeed, user rating is one of the most important features of

App advertisement. A higher rated App may attract more users to download and can also be ranked higher in the leader board. Thus, rating manipulation is also an important perspective of ranking fraud. Intuitively, if an App has ranking fraud in a leading session, the ratings during the time period of that leading session may have drastically changed patterns if seen from its previous historical ratings, which can be used for constructing rating based evidences. Rating to app is given by the user who downloaded it. 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 in number given as star to that App that is if 5 star given to the App is one in number among 100 users given other rating but not 5 star then it should be deleted and thus calculates rating score by summing all the ratings class collected and decision is taken on the basis of rating which scores high amongst all.

4. Review based evidences:

Including ratings, most of the App stores also allow users to write some textual comments as App reviews to submit to the developer. Such reviews can reflect the personal observations and usage understanding of breathing users for particular mobile Apps. Indeed, review management is one of the most important base of finding App ranking fraud. Specifically, before downloading or purchasing a new mobile App, users often first read its previous historical reviews to simplify their conclusion making and a mobile App includes more encouraging reviews may attract more users to download. Therefore, imposters often place counterfeit reviews in the leading sessions of a specific App in order to inflate the App downloads, and thus boost the App's ranking position in the leaderboard. Therefore, manipulation and detection of reviews is one way used over shady app developers to expertise the app. Hence reviews are used to detect the ranking fraud in Mobile App industry is the foremost viewpoint to find ranking fraud. On semantic analysis level review rechecking can be done to show the concluded review to user of app to make them easy to judge that app. As the Sentiment Analysis is a natural language processing task that deals with finding orientation of opinion in a piece of text with respect to a topic. To determine the semantic orientation of the sentences a dictionary based technique of the unsupervised approach is adopted. To determine the opinion words and their synonyms and antonyms WordNet is used as a dictionary. This module performs pre-processing of reviews and then performs sentiment analysis on pre-processed reviews. As the growing market of internet brought to the conclusion of product reviews as it made easy our decision about that product and as Internet is used by everyone the numbers of reviews that a product receives grow rapidly. To read all of comments is very time taking task for a potential customer and make a decision on whether to buy that product or not. Thus, mining this data about reviews, preprocessing that data, and classify them is an important task to make the reviews result corrected as shown below on stepwise proposing of such work: Gathering data for reviews from app store, and other sources: To determine the polarity of the sentences, based on aspects, large numbers of reviews are collected from the Web. There are lots of websites on the Internet where the large numbers of customer reviews are available. Amazon website (www.amazon.com) and also play stores like google play are used to collect the reviews. Preprocessing data to remove any missing entries (using filtering technique): To determine the semantic orientation of the sentences a dictionary based technique of the unsupervised approach is adopted. To determine the opinion words and their synonyms and antonyms WordNet is used as a dictionary; also, it plays a vital role in detecting any missing entries using filtering technique. Semantic matching for finding quality of review (Positive, Negative or Neutral): A large amount of reviews of users are collected on the Web that needs to be explored, analyze and organized for better decision making. Opinion Mining or Sentiment Analysis is a Natural Language Processing and Information Extraction task that identifies the user's views or opinions explained in the form of positive, negative or neutral comments and quotes underlying the text. Aspect based opinion mining is one of the level of Opinion mining that determines the aspect of the given reviews and classify the review for each feature. Semantic Matching: Algorithms and Implementation - Semantic Scholar.

Result

Fig 4.1 User Login

Fig 4.2 Global Anamoly Login

Fig 4.3 Local Anamoly Login

5.CONCLUSION

Here developed a rating fraud detection gadget for cell Apps. Specifically, right here first confirmed that rating fraud passed off in main classes and furnished a technique for mining main periods for every App from its historic rating records. Then, right here recognized rating primarily based evidences, ranking primarily based evidences and evaluation based totally evidences for detecting rating fraud. Moreover, right here proposed an optimization primarily based aggregation technique to combine all the evidences for evaluating the credibility of main periods from cellular Apps. A special standpoint of this method is that all the evidences can be modeled by using statistical speculation tests, as a consequence it is convenient to be prolonged with different evidences from area understanding to realize rating fraud. Finally, right here validate the proposed machine with large experiments on real-world App records gathered from the App store. Experimental effects confirmed the effectiveness of the proposed approach.

REFERENCES

[1] (2014). [Online]. Available: http://en.wikipedia.org/wiki/ cohen's_kappa
[2] (2014). [Online]. Available: http://en.wikipedia.org/wiki/ information_retrieval
[3] (2012). [Online]. Available: https://developer.apple.com/news/ index.php?id=02062012a
[4] (2012). [Online]. Available: http://venturebeat.com/2012/07/03/ apples-crackdown-on-app-ranking-manipulation/
[5] (2012). [Online]. Available: http://www.ibtimes.com/applethreatens- crackdown-biggest-app-store-ranking-fra ud-406764
[6] (2012). [Online]. Available: http://www.lextek.com/manuals/ onix/index.html
[7] (2012). [Online]. Available: http://www.ling.gu.se/lager/ mogul/porter-stemmer.
[8] L. Azzopardi, M. Girolami, and K. V. Risjbergen, "Investigating the relationship between language model perplexity and ir precision- recall measures," in Proc. 26th Int. Conf. Res. Develop. Inform. Retrieval, 2003, pp. 369–370.
[9] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," J. Mach. Learn. Res., pp. 993–1022, 2003.
[10] 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.
[11] 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.
[12] T. L. Griffiths and M. Steyvers, "Finding scientific topics," Proc. Nat. Acad. Sci. USA, vol. 101, pp. 5228–5235, 2004.
[13] 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.
[14] N. Jindal and B. Liu, "Opinion spam and analysis," in Proc. Int. Conf. Web Search Data Mining, 2008, pp. 219–230.