INTERNATIONAL JOURNAL OF ADVANCED RESEARCH IN COMPUTER SCIENCE AND ENGINEERING TECHNOLOGIES

ISSN 2454-9924 Volume: 1 Issue: 1(2015)

Mitigating Web Spam Taxonomy for Mobile App using Link Pruning and Reweighting Algorithm

V. Suriya #1, R. Mohan *1

Mailam Engineering College, Mailam #1 *1

Suriya.mangai@gmail.com #1

Abstract - In mobile application the fraud activities are widely spread and produce data loss. Ranking fraud in the mobile App market refers to fraudulent or deceptive activities which have a purpose of crash up the Apps in the popularity list. In real, it becomes more and more frequent for App developers to use covered means, such as inflating their Apps' sales or posting phony App ratings, to commit ranking fraud. In existing methods the prevention of fraud activities are not clearly determined ever. In this proposed system, we provide a rounded view of ranking fraud and web spam to detect and reduce the ranking fraud in mobile Apps systems. In this project, we provides link cutting and reweighting algorithm to find the web spam analysis and provide a broad coverage of various web spam forms. Using above algorithm Link spam and Click spam are determined. Link spam Adding links that point to the spammer's web site increases the page rankings for the site in the App Store. Similarly click spam, clicking ad banners without any motivation of purchasing the product.

Keywords – Mobile application, Fraud, Prevention, Spam, Rating

1 Introduction

The number of mobile Apps has grown at a breathtaking rate over the past few years. For example, as of the end of April 2013, there are more than 1.6 million Apps at Apple's App store and Google Play. To stimulate the development of mobile Apps, many App stores launched daily App leaderboards, which demonstrate the chart rankings of most popular Apps. Indeed, the App leaderboard is one of the most important ways for promoting mobile Apps. A higher rank on the leaderboard usually leads to a huge number of downloads and million dollars in revenue. Therefore, App developers tend to explore various ways such as advertising campaigns to promote their Apps in order to have their Apps ranked as high as possible in such App leaderboards. However, as a recent trend, instead of relying on traditional marketing solutions, shady App developers resort to

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH IN COMPUTER SCIENCE AND ENGINEERING TECHNOLOGIES

ISSN 2454-9924 Volume: 1 Issue: 1(2015)

some fraudulent means to deliberately boost their Apps and eventually manipulate the chart rankings on an App store. This is usually implemented by "Spam" inflate the App downloads, ratings and reviews in a very short time. Spam pervades any information system, be it e-mail or web, social, blog or reviews platform. In the literature, while there is some related work, such as web ranking spam detection, online review spam detection, and mobile App recommendation the problem of detecting ranking fraud for mobile Apps is still underexplored. To fill this crucial void, in this project, we propose to develop a ranking fraud detection system for mobile Apps. Generally speaking, web spam manifests itself as a web content generated deliberately for the purpose of triggering unjustifiably favorable relevance or importance of some web page or pages. It is worth mentioning that the necessity of dealing with the malicious content in a corpus is a key distinctive feature of adversarial information retrieval in comparison with the traditional information retrieval, where algorithms operate on a clean benchmark data set or in an intranet of a corporation. Daily App leaderboard became a de facto place to start information acquisition of mobile App. Though due to web spam phenomenon, search results are not always as good as desired. Moreover, spam evolves that makes the problem of providing high quality search even more challenging. Over the last decade research on adversarial information retrieval has gained a lot of interest both from academia and industry. In this project we present a holistic view of web spam detection for mobile App. When apps are downloaded from app store, Link spam and Click spam are detected using link pruning and reweighting algorithm. Link spam and click spam are comes under web spam. During app download the rating will be automatically incremented if the app is not affected by any spam (link and click spam), otherwise the rating standard with previous one. A spammer creates a page which looks absolutely innocent and may be even authoritative (though it is much more expensive), but links to the spammer's target pages. In this case an organically aggregated PageRank (authority) score is propagated further to target pages and allows them to be ranked higher. More aggressive form of a link spam schema is hijacking, when spammers first hack a reputable website and then use it as a part of their link farm. Spammers can also collude by participating

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH IN COMPUTER SCIENCE AND ENGINEERING TECHNOLOGIES

ISSN 2454-9924 Volume: 1 Issue: 1(2015)

in link exchange schemes in order to achieve higher scale, higher in-link counts, or other goals. We also consider redirection as an instant type of link spam. Here the spamming scheme works as follows. First, a link spam page achieves high ranking in a user review page by boosting techniques. But when the page is requested by a user, they don't actually see it; they get redirected to a target page. There are various ways to achieve redirection. The easiest approach is to set a page refresh time to zero and initialize a refresh URL attribute with a URL of a target page. More sophisticated approach is to use page level scripts that aren't usually executed by crawlers and hence more effective from spammers point of view. Since web use click stream data as an implicit feedback to tune ranking functions, spammers are eager to generate fraudulent clicks with the intention to bias those functions towards their websites. To achieve this goal spammers submit queries to a search engine and then click on links pointing to their target pages. To hide anomalous behavior they deploy click scripts on multiple machines or even in large botnets. The other incentive of spammers to generate fraudulent clicks comes from online advertising. In this case, in reverse, spammers click on ads of competitors in order to decrease their budgets, make them zero, and place the ads on the same spot. The App leaderboard is one of the most important ways for promoting mobile Apps. A higher rank App on the leaderboard usually leads to a huge number of downloads and million dollars in revenue. So, some fraudulent had happen to boost their Apps and eventually manipulate the chart rankings on an App store. In this project, we developed a ranking fraud detection system for mobile Apps. When apps are downloaded from app store, "Spam" inflate the App downloads, ratings and reviews in a very short time. Spam pervades any information system, be it e-mail or web, social, blog or reviews platform. Different types of spams are available in web. In this, we detect and overcome "Link spam" and "Click spam" using link pruning and reweighting algorithm.

2 Related Work

App stores launched daily App leaderboards to stimulate the development of mobile Apps. The App leaderboard is one of the most important ways for promoting mobile Apps. A higher rank App on the

ISSN 2454-9924 Volume: 1 Issue: 1(2015)

leaderboard usually leads to a huge number of downloads. Therefore, App developers advertising campaigns to promote their Apps in order to have their Apps ranked as high as possible in such App leaderboards. Instead of relying on traditional marketing solutions shady App developers do some fraudulent using bot farms to boost their Apps chart ranking on App store. Web ranking spam detection, online review spam detection, and mobile App recommendation are still under explored. Moreover, spam evolves that makes the problem of providing high quality search even more challenging. Here it provides some drawbacks are, It is difficult to detect when fraud happens, It is difficult to manually label ranking fraud for each App, Is not easy to identify and confirm ranking fraud, Search results are not always good, Problem of providing high quality search, Web spam is not detect.

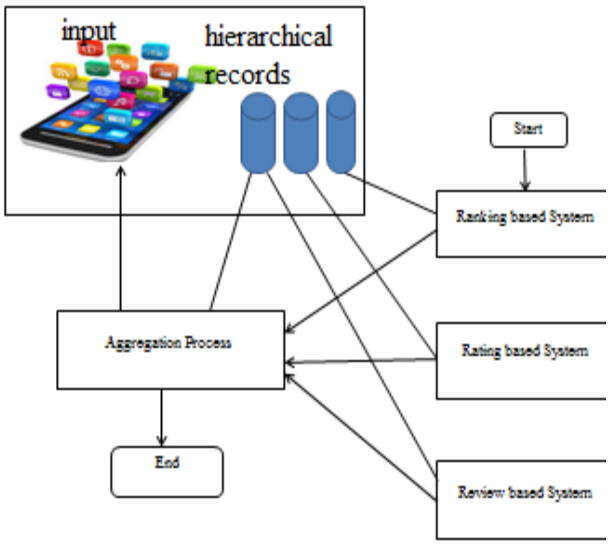
2.1 Proposed work

Apple's App store and Google Play became a de facto place to search and download Mobile Apps Store. Though due to web spam phenomenon, search results are not always as good as desired. So, we had to detect and avoid the web spam taxonomy. In our proposed work, we present a systematic review of web spam detection techniques with the focus on algorithms and underlying principles. Link spam and Click spam both are web spam discussed in our proposed work. Link spam Adding links that point to the spammer's web site increases the page rankings for the site in the App Store. Similarly click spam, clicking ad banners without any intention of purchasing the product. Clicking the ads countless times can make dishonest rankings in Mobile App Store. Link pruning and reweighting algorithms are used here to detect and avoid the web spam. Link pruning and reweighting algorithms detect the "nepotistic links", links that present for reasons rather than merit, for instance, navigational links on a website or links between pages in a link farm and also reports its resistance to fraudulent clicks. Here, it provides some benefits are, Automatic rating, Web spam detected, Link spam and click spam are detected, providing high quality search result.

3 Architecture

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH IN COMPUTER SCIENCE AND ENGINEERING TECHNOLOGIES

ISSN 2454-9924 Volume: 1 Issue: 1(2015)



3.1 Ranking based System

A leading session is composed of several leading events. There-fore, we should first analyze the basic characteristics of leading events for extracting fraud 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. Specifically, in each leading event, an App's ranking first increases to a peak position in the leaderboard (i.e., rising phase).

3.2 Rating based system

The ranking based evidences are useful for ranking fraud detection. However, sometimes, it is not sufficient to only use ranking based evidences. For example, some Apps created by the famous developers, such as Game loft, may have some leading events with large values of u due to the developers' credibility and the "word-of-mouth" advertising effect. Moreover, some of the legal marketing services, such as "limited-time discount", may also result in significant ranking based evidences. To solve this issue, we also study how to extract fraud evidences from Apps' historical rating records.

3.3 Review Based System

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 perspectives of App ranking fraud. Specifically, before downloading or purchasing a new mobile App, users often first read its historical reviews to ease their decision making, and a mobile App contains more positive reviews may attract more users to download.

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH IN COMPUTER SCIENCE AND ENGINEERING TECHNOLOGIES

ISSN 2454-9924 Volume: 1 Issue: 1(2015)

4 Methodology

4.1 Link Pruning and Reweighting

Notices that PageRank score of pages that achieved high ranks by link-spamming techniques correlates with the damping factor c. Using this observation authors identify suspicious nodes, whose correlation is higher than a threshold and down weight outgoing links for them with some function proportional to correlation. They also prove that spammers can amplify PageRank score by at most 1c and experimentally show that even two-node collusion can yield a big PageRank amplification. Where they show that due to the power law distribution of PageRank, the increase in PageRank is negligible for top-ranked pages.

4.2 Detection of Spam

Interesting idea to prevent click spam is proposed personalized ranking functions, as being more robust, to prevent click fraud manipulation. We present a utility-based framework allowing judging when it is economically reasonable to hire spammers to promote a website. The performs experimental study demonstrating that personalized ranking is resistant to spammers manipulations and diminishes financial incentives of site owners to hire spammers. The work studies the robustness of the standard click-through-based ranking function construction process and also reports its resistance to fraudulent clicks.

4.3 Web Graph

We model the Web as a graph with vertices, representing web pages, and directed weighted edges, representing hyperlinks between pages. If a web page(pi) has multiple hyperlinks to a page(pj), we will collapse all these links into one edge. Self-loops aren't allowed. We denote a set of pages linked by a page pi as Out(pi) and a set of pages pointing to pi as In(pi). Finally, each edge can have an associated non-negative weight.

4.4 Page Ranking Design

PageRank uses link information to compute global importance scores for all pages on the web. The key underlying idea is that a link from a page pi to a page pj shows an endorsement or trust of page pi in page pj, and the algorithm follows the repeated improvement principle. The true score is computed as a convergence point of

ISSN 2454-9924 Volume: 1 Issue: 1(2015)

an iterative updating process. The most popular and simple way to introduce PageRank is a linear system formulation.

5 Conclusion

From this, Detection of fraud in Mobile Application using Ranking has been implemented. In this project, we developed a web spam detection system for mobile Apps. To draw a general picture of the web spam phenomenon, we first provide numeric estimates of spam on the Web, discuss how spam affects users rating for mobile apps, and motivate academic research. In our project, we present a systematic review of web spam detection techniques with the focus on algorithms and underlying principles. Link spam and Click spam both are web spam discussed in our work. According to this work, web spam detection research has gone through a few generations: starting from simple content based methods to approaches using sophisticated link mining and user behaviour mining techniques.

6 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-fraud-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] A. Klementiev, D. Roth, and K. Small, "An unsupervised learning algorithm for rank aggregation," in Proc. 18th Eur. Conf. Mach. Learn., 2007, pp. 616–623.

[9] A. Klementiev, D. Roth, and K. Small, "Unsupervised rank aggregation with

ISSN 2454-9924 Volume: 1 Issue: 1(2015)

distance-based models," in Proc. 25th Int. Conf. Mach.Learn., 2008, pp. 472–479.

[10] A. Klementiev, D. Roth, K. Small, and I. Titov, "Unsupervised rank aggregation with domain-specific expertise," in Proc. 21st Int. Joint Conf. Artif. Intell., 2009, pp. 1101–1106.

[11] A. Mukherjee, A. Kumar, B. Liu, J. Wang, M. Hsu, M. Castellanos, and R. Ghosh, "Spotting opinion spammers using behavioral footprints," in Proc. 19th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2013, pp. 632–640.

[12] 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.

[13] 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.

[14] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," J. Mach. Learn. Res., pp. 993–1022, 2003.

[15] E.-P. Lim, V.-A. Nguyen, N. Jindal, B. Liu, and H. W. Lauw, "Detecting product review spammers using rating behaviors," in Proc. 19th ACMInt. Conf. Inform. Knowl. Manage., 2010, pp. 939–948.

[16] 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.

[17] J. Kivinen and M. K. Warmuth, "Additive versus exponentiated gradient updates for linear prediction," in Proc. 27th Annu. ACM Symp. Theory Comput., 1995, pp. 209–218.

[18] 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.

[19] N. Jindal and B. Liu, "Opinion spam and analysis," in Proc. Int. Conf. Web Search Data Mining, 2008, pp. 219–230.

[20] T. L. Griffiths and M. Steyvers, "Finding scientific topics," Proc. Nat. Acad. Sci. USA, vol. 101, pp. 5228–5235, 2004.

[21] Y. Ge, H. Xiong, C. Liu, and Z.-H. Zhou, "A taxi driving fraud detection

ISSN 2454-9924 Volume: 1 Issue: 1(2015)

system," in Proc. IEEE 11th Int. Conf. Data Mining, 2011, pp. 181–190.

[22] Y.-T. Liu, T.-Y. Liu, T. Qin, Z.-M. Ma, and H. Li, "Supervised rank aggregation," in Proc. 16th Int. Conf. World Wide Web, 2007, pp. 481–490.