ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 6, June 2017

Implementation of Aggregated Approach of Extracting Evidences to Detect Ranking Fraud

[1]Anamika R. Ghate, [2] Prof. U. A. Mande

[1] PG Student, Department of Computer Engineering, SCOE, Pune, India

[2] Associate Professor, Department of Computer Engineering, SCOE, Pune, India

ABSTRACT: These days everybody is utilizing advanced cells. There is need of different applications to be introduced on advanced mobile phone. To download application advanced mobile phone client needs to visit play store, for example, Google Play Store, Apples store and so on. At the point when client visit play store then he can see the different application records. This rundown is based on the premise of advancement or advertisement. Client doesn't know about the application. So user looks at the list and downloads the applications. Some of the time it happens that the downloaded application won't work or not valuable. That implies it is fraud in mobile application list. To stay away from this fraud, we are making application in which we will list the applications. To list the application first we will locate the dynamic time of the application named as leading session. We are contributing three sorts of proofs: Ranking based confirmation, Rating based confirmation and Review based proof. Utilizing these three confirmations at long last we are ascertaining total. We assess our application with genuine information gathered frame play store for long time period.

KEYWORDS: Evidence Aggregation, Historical Ranking Records, Mobile Apps, Ranking Fraud Detection, Rating and Review.

I. INTRODUCTION

In the past few years the number of mobile apps grows in highly incremental manner. Apple's app store and Google play store contains many apps. Leader board is used to demonstrate the chart rankings of most popular apps. For the promotion of mobile apps App Leader board is used. The app having higher rank in leader board leads to huge amount of downloading. So, the app developer acquires the various ways to ranked high their apps in the leader board i.e. advertising. Some app developers used the fraudulent way to promote their apps. They can manipulate the chart rankings in the app store. Sudden increase the app downloads, ratings and reviews is implemented by using"bot farms" and "human water armies". According to article from VentureBeat[1], when an app is promoted by using fraudulent way the ranking is increased from 1800 to top 25 and more than 50,000 100,000 new users could be acquired in a couple of days. This ranking fraud impact on mobile app industry in very large concern.

The literature work concern about mobile app recommendation [6],[9], online review spam detection[2] and web ranking spam detection[8],[10]. To overcome the problem of ranking fraud proposed ranking fraud detection system. There are some challenges to achieve this. First is that ranking fraud is not always happen so we have to identify accurate timing. Second challenge is that the number of mobile apps is huge so manually ranking each and every app is difficult. The apps are ranked high only in their leading session which is a collection of leading events. So to identify the leading sessions of each app based on its historical ranking records proposed algorithm. When the mobile apps promoted by using fraudulent way, specific pattern is observed. Some evidences are extracted by comparing this pattern with normal apps. To extract only ranking based evidences is not sufficient so we can extract rating and review evidences also. We used evidence aggregation method for the collection of these three types of evidences. All the evidences are heterogeneous. So the aggregation of this evidences is quite challenging. All the evidence extracted are heterogeneous so the modeling of the evidences are important. For the modeling of this evidences statistical hypothesis test are used. For the detection of review evidences KMP algorithm is proposed [11]. The proposed framework is scalable and can be extended with other domain generated evidences for ranking fraud detection.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 6, June 2017

II. REVIEW OF LITERATURE

There are three categories in which the related work of this study is separated. The first category is about the web ranking spam detection. The web ranking spam means that any web page having unjustifiable importance [8]. Ntoulas et al. [8] introduce a methods for the detection of web ranking spam detection and studied the content based web ranking spam detection. Zhoubet al. [10] introduce the problem of unsupervised web spam. They introduce the detection methods by using term spam and link spam. Spirin and Han [10] presented the algorithms and principles which reported the survey on web spam detection. By using ranking principles of search engines the web ranking spam detection is overcome. But this is not related to the ranking fraud detection system. The second category is about the online review spam detection. Lim et al. [2] detected the specific pattern of review
spammers and identified the spam by using review patterns. Wu et al. [9] introduce the problem of hybrid shilling attack on rating data. They proposed the system which mainly focuses on product recommendation. Xi et al. [7] focus on the problem of singleton review spam detection. To overcome this problem the can detect the patterns in multiple review based time series. All the above techniques are used to detect rating and review spam from historical records but it is not capable to extract evidences from given leading session. The third category is about the problems introduced in app recommendation system. Yan and Chen [9] introduce the mobile app recommender system which used the preference matrix to recommend apps instead of using ratings. Shi and Ali [9] proposed the model which solve the sparcity problem and proposed a filtering model which is a content based named Eigenapp. Zhu et al. [6] proposed a context aware recommendation system. From the literature we discussed above nothing can discussed about the ranking fraud detection system. The disadvantage of literature is

- Happens Fraud in Web Application.
- Leader Board gets false rating surveys of client.

III. COMPARATIVE ANALYSIS

Lim et al.[2] introduce the detection of product review spammers using rating behaviors. it proposes the scheme of the detected spammers have more significant impact on ratings compared with the unhelpful reviewers. It is used to improve the accuracy of the current regression model. But the disadvantage is that it is not clear how much review spam exists in online product review. We can refer the scalability of the detection algorithm as well as some regularity of ranking fraud activities. Xiong et al.[3] introduce a taxi driving fraud detection system. It proposes a taxi driving fraud detection system, which is able to systematically investigate taxi driving fraud. In city taxis to collect a large amount of GPS traces under operational time Constraints. The disadvantage is that GPS traces provide nonparallel opportunities for us to uncover taxi driving fraud activities. Jindal et al.[4] develop opinion spam and analysis system which analyzes such spam activities and presents some novel techniques to detect them. The disadvantage is that it has been neglected so far is opinion spam or trust worthiness of online opinions. A. Klementiev et al.[5] develop an unsupervised learning algorithm for rank aggregation. It is a novel unsupervised learning algorithm for rank aggregation which returns a linear combination of the individual ranking functions based on the assumption of profitable ordering agreement between the rankers. But the disadvantage is that rank aggregation problem by specifying an optimization problem. Xie et al.[7] develop a review spam detection via temporal pattern discovery. It can discover that singleton review is a significant source of spam reviews and largely affects the ratings of online stores. But a general detection algorithm will produce patterns to the problem in this paper therefore increase false positive rate.

IV. PROPOSED SYSTEM MECHANISM

The objectives of the proposed system are as follows:

- To focus on ranking fraud of mobile apps.
- To focus on online review spam detection.
- To analyze web ranking spam detection.

Detecting ranking fraud of mobile apps is mainly detecting ranking fraud in leading sessions of mobile apps. There are two challenges to detect ranking fraud in leading sessions of mobile apps.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 6, June 2017

- Identify leading sessions for mobile apps:

There are some preliminaries to show various ways to mine leading sessions from their historical records. Preliminaries include identifying leading events and leading sessions from records.

- Mining leading sessions:

There are two main steps of mining leading sessions. Firstly we have to mine leading events of popular mobile apps and second we have to merge that all leading events. Collection of these leading events is called leading session. From leading sessions to detect ranking fraud we have to extract evidences. There are three types of evidences. These are ranking based evidences, rating based evidences and review based evidences.

A. Ranking based evidences

Whenever any app ranked high in malicious way then it can have a specific pattern in leading session. This pattern can have three types of phases i.e. raising phase, maintaining phase and recession phase. App suddenly increase its popularity to peak position (raising phase), app keeps its peak position to specific time period (maintaining phase) and decrease its popularity (recession phase). Compare this pattern of raising, maintaining and recession phase with the pattern of normal apps. We can extract ranking evidences by making comparison of this.

B. Rating based evidences

Whenever any app upload then user wants to download it. User gives rating to that downloaded app according to its service given to user. The app having high rating makes high ranking in the leaderboard. Rating is the main feature of advertising. The app having high rating may attract large number of users to download the app. Thus, rating is the important fact of ranking fraud. If the leading session having ranking fraud then it's pattern of apps ratings is also different from ratings of normal apps. we can identify rating based evidences by comparing patterns of ratings.

- Compare the number of ratings in leading sessions with historical records.
-

C. Review based evidences

Whenever any user used some app then he/she wants to give comments according to service given by the app. This is the important popularity information like ranking and rating. This comment of users suggests the quality of service of app to other users. This comments can gives usage experience of existing user. So, the manipulation of review is one of the important fact of ranking fraud. When any user wants to download any app then he/she read the comments given by the existing user. This can help for the decision making of new user. More number of positive comments or reviews can attract more users to download. So, the manipulator can post fake reviews in order to increase the download. The ranking position of app also increased. To overcome this review based fraud proposed two evidences. Review spammers always do duplicate comments. This is the main evidence. Review spammers do reviews related to the specific topic such as worth to play and very boring. By extracting this type of evidences from the leading session detection of review based ranking fraud would be possible.

- Compute average mutual similarity between reviews within leading session
- Manipulated reviews contain more positive topics.
- Early Time Frame.
- Maximum Number of Reviews

D. Evidence Aggregation

All three types of extracting evidences are heterogeneous. So, it is the challenged to aggregate this data to detect ranking fraud. In the literature aggregation methods are used such as score based model, Dampster-Shafer rules and permutation based models. But this methods are not proper for new apps. Because this methods mainly focus on global ranking of apps. Other methods used are supervised learning techniques. But this type of technique are not scalable to other domains. So, proposed un-supervised approach based on fraud similarity to extract evidences. The advantage of proposed system is that it is versatile and can be reached out with other domain produced evidences for ranking fraud detection.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 6, June 2017

V. SYSTEM ARCHITECTURE

Fig. 1 shows system architecture. It includes the ranking, rating and review based evidences which were extracted to detect ranking fraud.

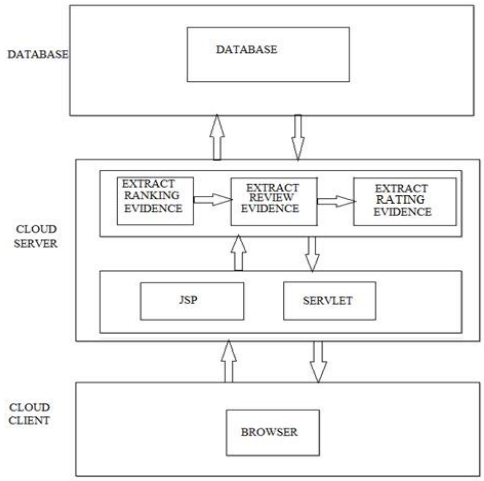

Fig.1. Overview of ranking fraud detection system

VI. ALGORITHMS USED

Mainly ranking fraud occur in the leading session. From this leading session we can extract evidences which can detect the malicious behavior of app that is ranking fraud.

KMP algorithm:

The working of KMP algorithm is similar to the naive algorithm. It can correlate the element with the neighboring element in order to 1 to n-m. It is used to detect the matching elements. The difference is that the KMP algorithm uses information extracted from partial matches of the pattern and text to glance over shifts that are guaranteed not to result in a match. Suppose that, starting with the pattern adjusted beneath text at the leftmost end, we repeatedly shift the pattern to the right and try to match it with the text.

We used KMP algorithm for the detection of review evidences. Suspicious reviews can contain more positive topics. To identify such patterns in reviews we used KMP. KMP is also used for the searching operation of apps by user in the system.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 6, June 2017

Algorithm 1 Knuth-Morris-Pratt Algorithm
Require: P, T

```
KMP-Matcher (T, P)
1    n ← length [T]
2    m ← length [P]
3    π ← Compute-Prefix (P)
4    j ← 0
5    for i from 1 to n
6         do while j > 0 and P [j + 1] ≠ T[i]
7              do j ← π[j]
8           if P [j + 1] = = T[i]
9              then j ← j + 1
10        if j = = m
11             then print "String found at", (i – m)
12                  j ← π[j]

Compute-Prefix (P)
1    m ← length [P]
2    π [1] ← 0
3    k ← 0
4    for j from 2 to m
5         do while k > 0 and P [k + 1] ≠ P[j]
6              do k ← π[k]
7           if P [k + 1] = = P[j]
8              then k ← k + 1
9           π [j] ← k
10   return π
```

Fig 2. KMP algorithm.

VII. EXPERIMENTAL SETUP

To analyze and study the performance of ranking fraud detection system we have to evaluate some experimental setup. Firstly we have to collect datasets for the experimental setup. These data set were collected from the Google App store. These sets were extracted from the Top Free 100 and Top Paid 100 apps. The experimental setup consists of top 50 most suspicious leading sessions which is top ranked, 50 uncertain sessions which is middle ranked and 50 normal leading sessions which is bottom ranked from the data set. Then we have to select all the selected sessions.

VIII. PROPOSED RESULT

The experimental data were collected from leader boards of Google's app store. Data will be top free 100 apps and top paid 100 apps. Data set will be consists of rating and review also. We can observe the behavior on the basis of popularity information of apps.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 6, June 2017

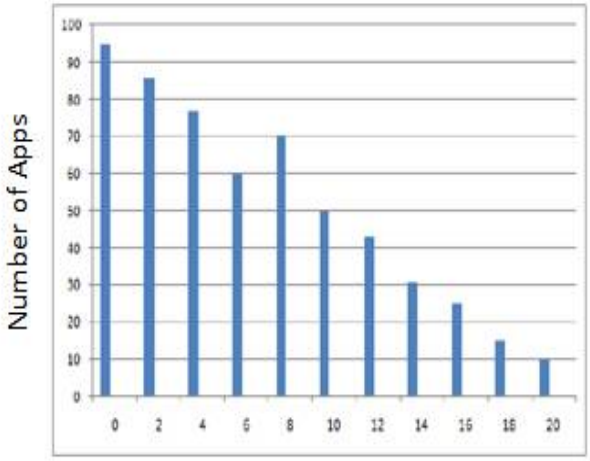

Fig. 3. Ranking behavior of leading sessions of top free 100 apps for data set.

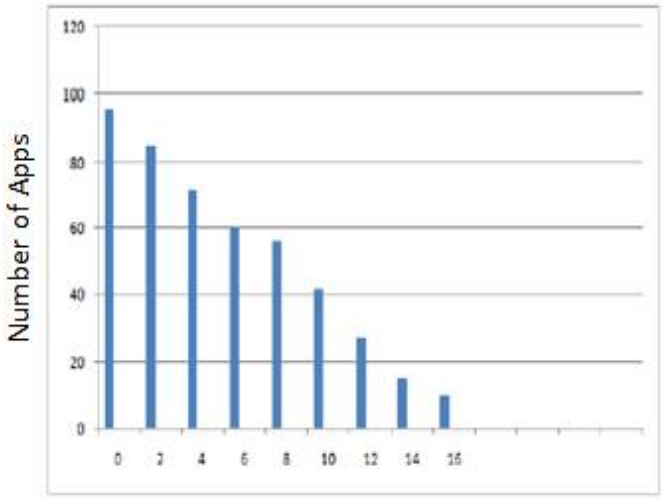

Fig.4. Rating behavior of leading sessions of top free 100 apps from data set

The above figures 3 and 4 will be the different number of leading sessions on the basis of distribution of number of apps with respect to their rankings and ratings in the data set. From the above figures, we can observe that the number of apps which is ranked high is less than the apps which is ranked low. The above figures 3 and 4 will be the different number of leading events on the basis of distribution of number of apps with respect to their ratings in the data set. From the

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 6, June 2017

TABLE I
CHARACTERISTICS OF DATA

	Top Free Apps	Top Paid Apps
App Num	9,784	5,261
Ranking Num	285,900	285,900
Rating Num	14,912,459	4,561,943

figures we conclude that the number of apps high is low on the basis of ratings. Table 1 shows the detailed data characteristics.

A. Mining Leading Session

From the above figures we can see that only a few apps have many leading events and leading sessions. We can find only a few leading sessions contain many leading events. This also validate the ranking evidence. We develop some baselines to validate the effectiveness of Evidence Aggregation based Ranking Fraud Detection. The first baseline stands for ranking evidence based ranking fraud detection, which calculates ranking fraud for each leading session by only using ranking based evidences. These three evidences are aggregated by specific approach. The second baseline stands for rating evidence based ranking fraud detection, which calculates the ranking fraud for each leading session by only using rating based evidences. Then these evidences are aggregated. The third baseline stands for review evidence based ranking fraud detection, which estimates the ranking fraud for each leading session by only using review based evidences. Then these evidences aggregate by using specific approach.

B. Overall Performance

Systems are consistently outperforms baselines and the improvements are more significant for smaller thresholds. The results validates the effectiveness of evidence aggregation based
framework for detecting ranking fraud. Baseline indicates that rank based aggregation for integrating fraud evidences. Leveraging three kinds of evidences is more effective than only using one type of evidences, even if without evidence aggregation. Review manipulation does not directly affect the chart rankings of apps, but may increase the possibility of inflating app downloads and ratings. Result found that the review based evidences could always improve the detection performances while being used together with other evidences. Only leveraging single kind of evidences for fraud detection cannot obtain good performance means finding such suspicious apps in high positions.

C. Efficiency of Approach

The computational cost of this approach mainly comes from the task of extracting three kinds of fraud evidences for the given leading sessions. The fraud signatures of existing leading sessions can also be mined in advance and stored in the server. The process of extracting evidences for each leading session will be very fast.

D. Effectiveness of Algorithm

We use KMP in our system. KMP is pattern matching algorithm. This algorithm reduces the number of comparisons and parallelization improves the time efficiency. This algorithm achieves a better result as compared to the multithreaded version of the algorithm where again by text dividing, the parallelization is achieved. This algorithm is efficient in worst case. KMP matching algorithm is a substantial improvement for the deficiencies of the other algorithm. Whenever the match is unsuccessful the pointer doesn't go back to the first position, it returns to the proper place. It is the key of KMP algorithm. In our system, for the detection of evidences we can match the patterns of fraudless app to the malicious app. In the existing system pattern matching is done by statistical hypothesis test. This test can somewhat manually done. In our proposed system, we used KMP algorithm for pattern matching. This can enhance the overall performance of system. Matching can be done automatic and process is time consuming. So the proposed system is more efficient than existing. The fraud detection done is in less time as compared to existing. Process is in faster state. Specifically KMP is used in review matching process. It is used to compare the reviews of different users of specific app. By comparing it, it can detect the duplicate reviews and prevent duplication. So, whenever the user validate the reviews he/she cannot see the fake

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 6, June 2017

reviews. By doing this, KMP algorithm can increases the efficiency of detection method of ranking fraud. Also user can validate and verify apps by observing ranking graph. Leading event can be estimated by using number of downloads per day. Ranking can be calculated by aggregating ranking, rating and review of each app.

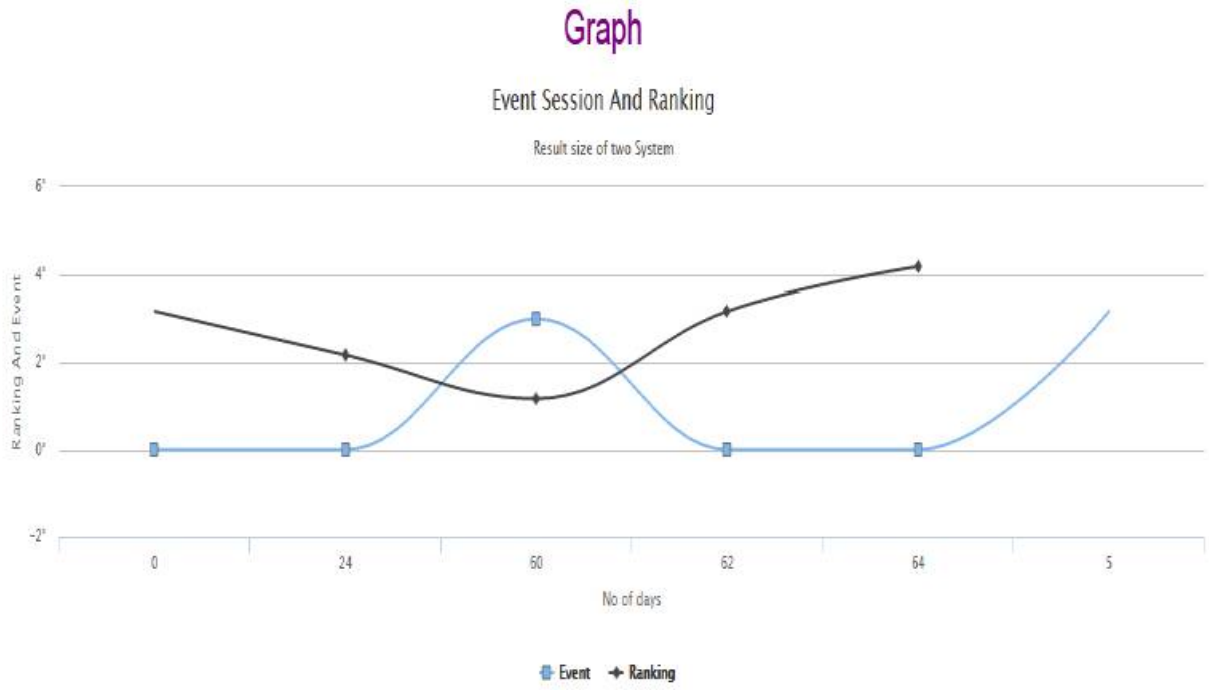

Fig. 5 Ranking record of app.

The figure 5 shows the ranking record of Angry Bird app. This type of graph is generated for each and every app in the dataset. The blue graph shows the events of app and black graph shows the ranking of app. The final result of this system shows the list of apps with the percentage of fraud detects in it.

IX. CONCLUSION

Work can built a ranking fraud detection system for mobile Apps. To built such a system , first showed that ranking fraud detected in leading sessions and provided a algorithm for mining leading sessions for each App from its historical ranking records. Then, extract ranking based evidences, rating based evidences and review based evidences from the historical records to detect ranking fraud. To overcome this fraud proposed an optimization based aggregation method to integrate all the evidences for estimating the credibility of leading sessions from mobile Apps. A different perspective of this approach is that all the evidences can be modeled by statistical hypothesis tests, thus it is easy to be enlarged with other evidences from domain knowledge to detect ranking fraud. Finally, validate the proposed system with extensive experiments on real-world App data collected from the Google Play store.

REFERENCES

[1] Hengshu Zhu, Hui Xiong, Yong Ge, Enhong Chen, “Discovery of ranking frauds for mobile apps,”, IEEE transactions on knowledge and data engineering, vol. 27, no. 1, january 2015.
[2] Ee-Peng Lim, Viet-An Nguyen,Nitin Jindal, Bing Liu, Hady W. Lauw, “Detecting Product Review Spammers using Rating Behaviors”, CIKM10, October 2630, 2010.
[3] Y. Ge, H. Xiong, C. Liu, and Z.-H. Zhou, “taxi driving fraud detection system”, in Proc. IEEE 11th Int. Conf. Data Mining, 2011, pp. 181190.
[4] N. Jindal and B. Liu, “Opinion spam and analysis,”in Proc. Int. Conf. Web Search Data Mining, 2008, pp. 219230.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 6, June 2017

[5] A. Klementiev, D. Roth, and K. Small, "An unsupervised learning algorithm for rank aggregation,",in Proc. 18th Eur. Conf. Mach. Learn., 2007, pp. 616623.
[6] 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. 12121217.
[7] 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. 823831.
[8] 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. 8392.
[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. 204212.
[10] 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.
[11] Zachary K. Baker, and Viktor K. Prasanna,"A Computationally Efficient Engine for Flexible Intrusion Detection", IEEE TRANSACTIONS ON VERY LARGE SCALE INTEGRATION (VLSI) SYSTEMS, VOL. 13, NO. 10, OCTOBER 2005