International Innovative Research Journal of Engineering and Technology
ISSN: 2456-1983 Vol: 3 No: 3 March 2018

Counterfeit Review Rating and Ranking Fraud Detection

Emmanuel O.C. Mkpojiogu
Department of Computer and Information Technology, Veritas University Abuja, Nigeria

Abstract: Nowadays the online market has been a recent trend all over the world. They provide all products and services to all people. These products and services are greatly trusted with the help of reviews on sites. From software applications to hardware products everything has their own review. Some of the product creators modify the ratings by giving false review to their product. The people who see such review gets confused and uses them and feel unworthy about the product sometimes about the online market. By giving the false review the product takes top place in ranking. But there is a great demand for getting rid from these review frauds. Some solutions focus on the classification of opinions using the natural language processing and data mining. We have a new way to get rid from these review frauds. We investigate the evidences by modelling the products and services based on the behaviour. By collection the information and by the analysis we and find the review frauds. We use the OLAP data aggregation for online reporting mechanism for information process. All feedbacks from user are collected. User can able to differentiate the fake and original products and services. In addition, the user interest can also be recorded in this method. This is employed by data mining and artificial intelligence to find the fake profiles and their commands.

Keywords: Fraud review detection, fake rating, user interest.

1. INTRODUCTION

As the technology increases tremendously the products also developed using those technologies. Those are available at the local store. An additional market called online market is now emerging as the products are delivered at our home step. Not only the demand for the product the demand for services also increased nowadays. There are available experts whose services can be also get those online market. Every product and service has its own purpose. Based on the user demand the developers develop the different products for the users. Due to demand of user recommendations several products are created for same purpose.

For an example the basic model cell phones are replaced by the smart phones which has the basic and the additional features. As the demand for products and services increases there were many product developers and service providers who provide same solution. Anyhow between a battle one may win other need to lose. Same happens effectiveness of the product and look and other specifications are the important weapons for wining this war. The judgment is provided by the user ratings and reviews.

2. DETECTING FRAUDS AND EVIDENCES

There were many existing ways to detect the review frauds. These traditional ways are complex. These time-consuming method investigate the dissimilar domains of knowledge. Fraud perform many instances using similar method. Fraud is a different method of performing crime which needs a special data analysis for finding those. This method uses the areas of knowledge discovery in database, data mining, statistics and machine learning.

A. Evidence in Ranking

The analysis of the ranking and finding whether any patterns are there like rising phase, maintaining phase or reducing phase. Usually when the product gets popular then its rank will increase this is called rising phase. Then it will maintain its state in the market this phase is maintain phase. The period when its popularity reduces is the recession phase.

The ranking based evidences provide a good information but does not provide the better accurate results. Because due to the benchmarking of a company their product may be at peak during the release of that product due to the advertisement.

Later they may fall down or stable on the place based on the functionality of that specific product or service. Some discounts for specific time also increases their rank.

B. Evidences in Rating

So for the better accuracy we also find the rating of a product and services. We compare the patterns of ranking

with the patterns of rating for finding whether a product has an anomaly in patterns. The fraudulent product or service have increasing rating over a short period of time. This can be confirmed by finding the average rating on every day.

C. Evidences in Reviews

Also the users may provide an additional text format commands on those products called as review. They can provide their personal opinion on the product or service. This also provide an important role in the product and service ranking fraud. In general opinion the product with more positive review is used more likely than other product. Therefore, some providers post fake review to attract most users to use their product which may have an impact in their rank.

3. PROPOSAL SURVEY

A. Latent Dirichlet Allocation [1]

In this paper the proposer proposes the collection of discrete data by generative representation by describing the Latent Dirichlet Allocation (LDA). It is a hierarchical model which has three layers collecting the finite mixture of fundamentals. Each has the infinite mixture of topics. The document is explicitly modelled using the probabilities in the text modelling framework. EM algorithm and inference approximation are used by authors for the empirical Bayes parameter evaluation. The results are reported in the collaborative filtering, text classification with comparison to some unigrams and LSI representation.

B. Opinion Spam and Analysis [2]

The paper focus on the customer review which is the main focus of the consumers and producers. Nowadays the web spam and email spam are studied and have enormous solutions to that problem. But for our topic we discuss about the opinion spam which is hard to find and very less solutions are proposed which also does not provide a better solution. It needs very complex detection methods as they may vary from user to user.

C. An Unsupervised Learning Algorithm for Rank Aggregation [3]

The review of the user may vary from the user to user and it does not depend on the common thing is the great difficulty found in the data retrieval. There comes some method called the rank aggregation by merging the rank of various products or services. This proposed Unsupervised Learning Algorithm for Rank Aggregation is used for its strength in the data fusion. This is clearly seen over the ad-hoc retrieval system.

D. Supervised Rank Aggregation [4]

It is focussed on the rank aggregation which is formerly done by the unsupervised learning. The accuracy of rank aggregation is achieved in this method as it follows the supervised methodology. They label the data for getting the supervision. It is hard to get those results as it follows the Markov chain. But the results show that the accuracy is higher than the older unsupervised methods.

E. Spotting Opinion Spammers using Behavioural Footprints [5]

The individuals and the organizations are able to post their review on the sites which are accessible by the other users. This paper is proposed to find the spammers in a new different angle. The footprint of the users is tracked and their behavioural knowledge is obtained by this method. By this the observation create a two different cluster one is of spammers and other are non-spammers. This is done by general observation the spammers and non-spammers behavioural footprints are diverge in nature. Heuristics and ad-hoc labels are used for the spam detection in existing methods. This also has the way to find the results of unsupervised opinion without use of physical data to supervised model.

F. Survey on Web Spam Detection: Principles and Algorithms [6]

In this the new way for finding the spammers are created and this method classifies the existing methods into categories. The criteria for the classification is based on type of information. Context based methods, methods based on non-traditional data. The non-traditional data may be user behaviour, clicks, sessions and link-based category. Further the link based category is sub divided. The numerical data for finding spam is proposed in this methodology. It uses the algorithms and some principles for finding the spammers and also to create an awareness to simulate further research.

G. A Semi-Supervised Hybrid Shilling Attack Detector for Trustworthy Product Recommendation [7]

The product is placed in the top of the review based on user review and their ratings. If this product is worthy, then they are placed in the top of the recommendation list by the recommender manipulation system. But shilling attack apply some fake ratings to get their desired product in the recommendation list. To wrap these problem, this paper uses the Hybrid Shilling Attack Detector (HySAD). This semi-supervised model differentiates the average and random filler model attackers from the normal users. The effectiveness of HySAD is found by the experimenting this method in various sites like Movie-Lens and Netflix. It also improves the accuracy of the collaborative filtering based recommender system to get more in-depth details about the spammers.

4. PROBLEM DEFINITION

Everyday new products are launching in the market. The ratings and reviews are changing on these markets every day. There is no proof or statement that the review fraud happens in every product. It may happen sometimes it may not even. The hard part of the thing is we need find those thing whether the fraud has been taken place or not. Therefore, to solve this we need find the anomalies to find these rating and review frauds. There are more number of products available online. For manual detection it takes enormous time resource and careful means. So we are moving to the efficient way that the machines automatically find the ranking fraud.

5. EXISTING SYSTEM

The existing system of spam detection has several methods. They find the spammers in the web by using the spam detection method. They are found by the ranking on search engines and also by analysing the behaviour of the spammers. The analysis mainly focuses the rank aggregation. If product launched then based on the popularity of the product it goes to increasing phase or the maintaining phase after a time interval it will be in maintaing phase or moves to decreasing phase. If it remains in the increasing phase then or if it makes sudden increasing in phase then it may predict it as the product has some fake reviews. A sample graph to show how the ranking for applications is done is shown in fig1.

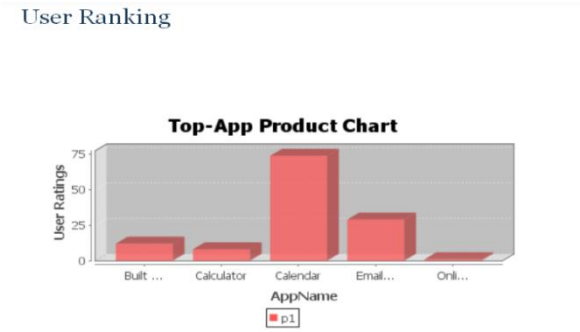

Figure 1. Graph for user ranking drawn using rating

A. Disadvantages of existing system

- The main disadvantage is the focus on the classification, because the different product may have different classification of categories.
- If the provider provides an offer during events the ranking may increase which will be detected as fake reviewed product.
- In earlier way, the fraud is not able to find as the rank is the only measure to find the fraud in the product or services.

6. PROPOSED SYSTEM

The system is presented to detect the review and rating frauds for the products and the services. Previous evidences are taken for the further investigation. But only these evidences cannot make the results accurate. The user is trying to get a product or the service. If he has only one option, he has to go for it. If there available several options for selection, he wishes to get the excellent product (say) (He may also choose the product based on the price also.). On consideration with the quality of product the user cannot get all the products and try all those and find the better quality of product which may be in great loss of time and money. The user reads the user review and rating on products and he then chooses them based on these reviews. If, there available false reviews then the user may get confused about the product. The use case diagram can help us to easily identify the actors in the system and their role. The use case diagram for the fraud detection system is in Fig2. If, the reviews are positive for that particular product then it tempts the user to use those resource. Therefore, we need more evidences to prove the false review frauds. Based on the historical records and the new evidences the new record has been generated which is more efficient. The reviews are aggregated using the method called data aggregation. The architectural model of the system specifically for mobile application is shown in Fig 3.

International Innovative Research Journal of Engineering and Technology
ISSN: 2456-1983 Vol: 3 No: 3 March 2018

From this methodology, the user interest is also measured. Online Analytic Processing (OLAP) is the method used for mining the results. The information in the OLAP is processed using the online reporting mechanism.

Both the positive and the negative feedback from the users are collected and saved in this system. First the user uses the product and then he gives the review of the product based on this review the original and false product can be identified. Now for getting the better results the apriori and the k-means clustering data mining algorithms needs to be used. After getting the inputs stored for analysis its classified using these two algorithms. k-means cluster algorithm groups the user review and the apriori algorithm then finds the larger dataset in those clusters and get those exact results. Now after the analysis, the fake review can be found and reported to the administrator. Now he can remove the fake reviews if he wants.

Here comes the final and the important part of the proposal. By the k-means and apriori algorithm the results are varied and found out. After the results found we then use the artificial intelligence programming which can find the fake reviews and report it to the administrator. In this the apriori algorithms result (frequent items) is taken for the comparison with the user profile which has the user interest

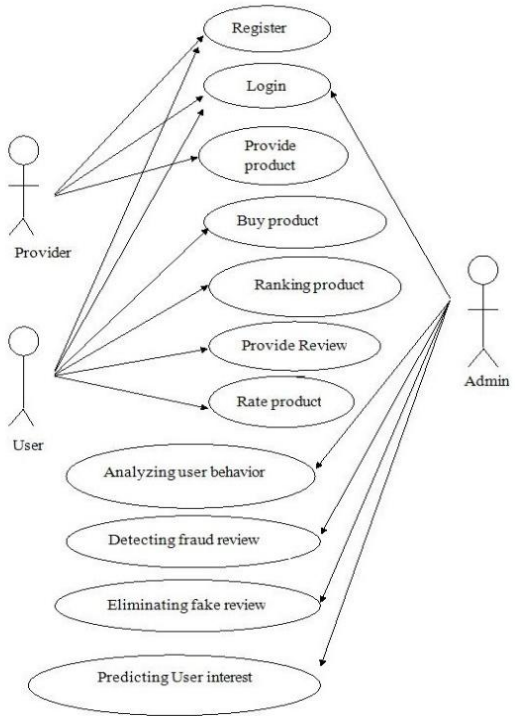

Figure 2. Use case diagram of the system.

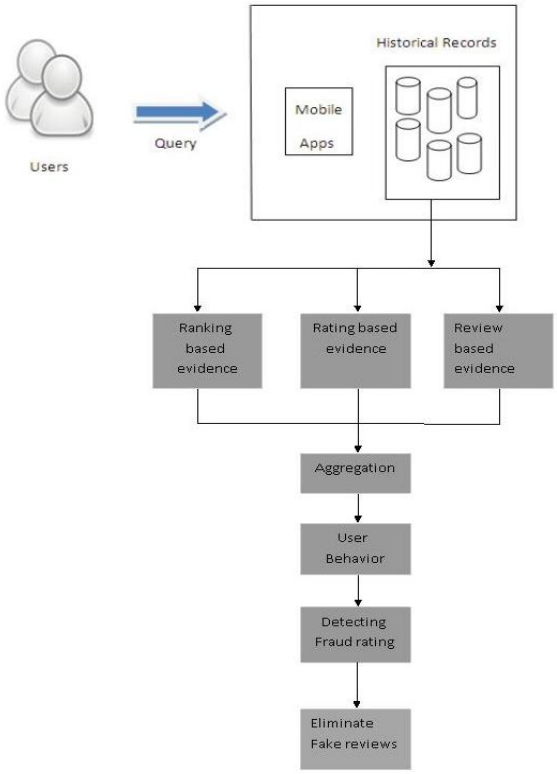

Figure 3. Architecture of system

Based on the study with the user relationship and the product the reviews are analysed. In the user profile his total reviews and rating history is saved which is very helpful for finding the fake reviews. If the user provides more number of fake reviews even after providing the warning to the user, his account will be reported to the administrator.

A. Advantages of proposed system

- Efficiency of the results are high compared to existing system.
- The review and rating frauds are analysed in an effective way.
- It completely avoids the ranking frauds by finding them.
- The complete online market can use this system without any constrain.

B. K-means clustering algorithm

K-means clustering algorithm is used to find the clusters in the given dataset. It is a data analysis algorithm for classifying the given dataset. The cluster which is the groups of the dataset. K-means algorithm is a heuristics algorithm. We use the algorithm to classify the user

review as different groups or clusters. Further operations are done with these clusters as input.

C. Apriori algorithm

In this apriori algorithm, the clustered input is taken and the results are provided to next step. The algorithm will return the frequent occurring dataset in the given input. This algorithm is very effective in proving our result. The repeated transactions are found using this effective algorithm. By using the bottom-up approach the algorithm will return the accurate results.

7. IMPLEMENTATION

The important stage in any proposal is the implementation. The theory is practically implemented. The implementation will solve the existing problem and has a better solution for the problem. The prototype of the review table is shown in the Table1. There are several parameters which can change the theoretical results and the practical results. The students who just learned them are strong in the theoretical but they lack the implementation and practical knowledge. The prototype of the provider details is provided in the table 2.

A committee is developed for the implementation co-ordination on individual organizations are appointed. The step of implementation begins with creating a plan on implementation. The plan carries the steps to follow and the complete manual of the project. It also holdsthe resources required for the solving this implementation.

Now after creating the plan it has to be followed. This stage shows the result of the system in efficient manner with proof. The system is build and needs to be tested before the implementation of the system. This testing phase is also important so the old errors can be eliminated.

Table1. Prototype of Review Table

Attributes	Data Types	Size	Constraints	Description
Product	Varchar	50	Not Null	Name of product
Comp	Varchar	50	Not Null	Name of company
Feedback	Varchar	50	Not Null	User feedback
Rank	Varchar	50	Not Null	User ranking
Review	Varchar	50	Not Null	User review

Table 2. Prototype of Provider Details

Attributes	Data Types	Size	Constraints	Description
Id	Varchar	20	Primary Key	Provider ID
Comp	Varchar	20	Not Null	Name of company
Username	Varchar	30	Not Null	Username of provider
Email	Varchar	50	Not Null	Email id of provider
password	Varchar	20	Not Null	Security password to login

8. CONCLUSION

Thus by proposing this system an effective ranking fraud detection is developed. Now the products and services can be only with the original reviews and ratings. Fake ratings can be eliminated by our system enabling the user to get the original reviews about the product. The ranking based evidences, rating based evidences and also the review based evidences are collected form the user and recorded for the analysis of the fake ranking detection. By this evidences the better fraud analysis is done. After the detection the fake reviews and ratings can be removed. Administrator also gets the users interest based on those evidences. Thus the fraud detection on any product and services are done with great efficiency.

REFERENCES

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

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

[3] Gleich D. F. 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] Ge Y, Xiong H., Liu C., and Z.-H. Zhou, "A taxi driving fraud detection system," in Proc. IEEE 11th Int. Conf. Data Mining, 2011, pp. 181–190.

[5] Heinrich.G, Parameter estimation for text analysis, "Univ. Leipzig, Leipzig, Germany, Tech. Rep.,

http://faculty.cs.byu.edu/~ringger/
CS601R/papers/Heinrich-GibbsLDA.pdf, 2008.

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

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

