Finding Fraud Rank in Spatial Database Using Reddit Ranking Algorithm

[1]Premkumar T, [2]Karthikeyan S.

[1]MSc (DSBA), [2]M.Sc., Ph.D.
[1]PG Scholar, [2]Assistant Professor
Department of Computer application,
Rathinam College of Arts and Science, Coimbatore, India.

***Abstract*: Detect fraud ranking persons in any online e-commerce application reviews play a very important role. Large part of the customers read reviews of products or stores before making the decision of what or from where to buy and whether to buy or not. As writing fake reviews comes with monetary gain, there has been a huge increase in deceptive opinion spam on online review websites. Basically fake review or fraudulent review or opinion spam is an untruthful review. Positive reviews of a target object may attract more customers and increase sales as well as negative review of a target object may lead to lesser demand and decrease in sales. Both are comes under fake review category only. These fraudulent reviews are deliberately written to trick potential customers in order to promote them. This work is aimed to identifying whether a review is fake or truthful one.**

In the present scenario, customers are more dependent on making decisions to buy products either on e-commerce sites or offline retail stores. Since, these reviews are game changers for success or failure in sales of a product, reviews are being manipulated for positive or negative opinions. Manipulated reviews can also be referred to as fake/fraudulent reviews or opinion spam or untruthful reviews. In today's digital world deceptive opinion spam has become a threat to both customers and companies. Distinguishing these fake reviews is an important and difficult task. These deceptive reviewers are often paid to write these reviews. As a result, it is a herculean task for an ordinary customer to differentiate fraudulent reviews from genuine ones, by looking at each review.

***Keywords:* Data Mining, Web Mining, Data Classification, (FFRS) - Reddit Ranking Algorithm, (RRA) - Finding Fraud ranking in spatial database**.

1.0 INTRODUCTION

The term web based E-commerce structure is identified with programming systems for internet business applications. They offer a domain for building online business applications rapidly. Web based business structures are sufficiently adaptable to adjust them to your particular necessities. As result, they are reasonable for building for all intents and purposes a wide range of online shops and internet business related (web) applications.

A web based business structure must

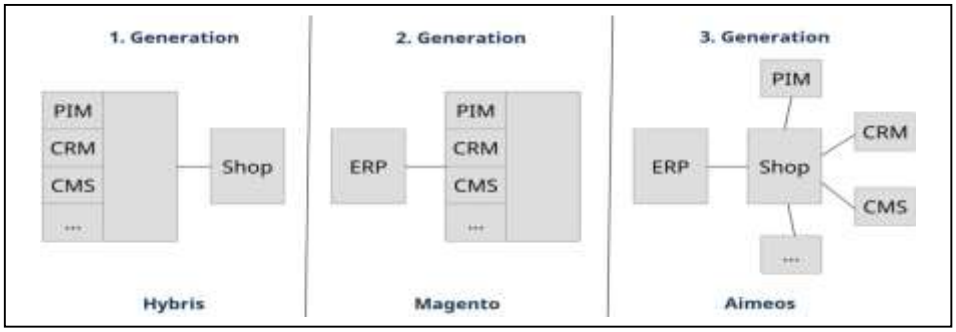
- ✓ IT permits supplanting all parts of the system code
- ✓ The restrict changes in the system code itself
- ✓ IT contains bootstrap code to begin the application
- ✓ And it be extensible by client composed code

Web based business systems ought to

- ✓ It characterize the general program stream
- ✓ Consist of reusable parts
- ✓ Be composed in useful areas

They give a general structure to web based business related applications. Moreover, they execute the general program stream (e.g.) how the checkout procedure functions. In opposition to solid shop frameworks, existing system stream can be stretched out as well as totally changed by your necessities. Since the start of (web) web based business around 1995, a ton has changed on the innovation side. The original of internet business frameworks advanced from existing ERP and related frameworks. This was trailed by the second era of independent shop frameworks in the vicinity of 2004 and 2008. Internet business structures are the most recent age of web based business frameworks and began around 2012. Hybrids, the shop framework possessed by SAP are one of the agents of the first era. It's unequivocally associated with the SAP ERP framework and Hybrids is predominantly a shop front-end for SAP. Client Relationship Management (CRM) and substance administration (CMS) apparatuses are accessible in the ERP framework however exceptionally constrained.

Fig 1.1: E-commerce Platforms



2.0 E-COMMERCE CLASSIFICATION

E-Commerce Classification E-commerce for SMEs can be generally classified into the following four categories:

- ✓ Business-to-Customer (B2C)
- ✓ Business-to-Business (B2B)
- ✓ Business-in-Business (B1B), and
- ✓ Consumer-to-Consumer (C2C)

Let us look at the first two quickly. These two, B2C and B2B, are the current primary initiatives in e-commerce being pursued by organizations today. Business-to-customer (B2C) e-commerce is concerned with commerce associated with the individual end consumer. This type of e-commerce is typically characterized by high volume, low value transactions across a broad customer base. Business-to-customer e-commerce is generally perceived, or marketed, to be Internet applications enabling customers to purchase products or services using a Web based application. Other technologies are available that can satisfy B2C e-commerce. Some of these technologies are:

- ✓ facsimile machine – in a variety of configurations and ways,
- ✓ computer to computer such as EDI,
- ✓ electronic publishing – publisher and subscriber,
- ✓ electronic funds transfer at point of sale (EFTPOS),

Business-to-business (B2B) online business is between associations. This might be between SMEs, expansive ventures, or amongst SMEs and huge undertakings. This kind of online business can be portrayed by low volume, high esteem exchanges over a thin client base. A portion of the web based business innovations utilized as a part of B2C online business can be similarly connected to B2B internet business (e.g. copy, IVR, and electronic inventories and indexes). Be that as it may, there are some particular advances that are suitable for B2B online business, for example, EDI and money related exchanges. Intra-business web based business (B1B) is frequently cited as a third class of web based business.

Intra business web based business is trade inside an association. This may appear as a work process application that backings a business exchange through the different specialty units of an association. Additionally, vast associations may in their own correct lead business exchanges between their different specialty units. For instance, the IT division may charge different specialty units for IT administrations. Intra-business online business is essentially material to vast undertakings. SMEs have excessively couple of workers, making it impossible to financially maintain intra-business IT&T applications for online business. Customer to purchaser (C2C) is a territory that is just barely starting to be taken note. This includes two purchasers getting together by means of an electronic means, for example, email, and making an understanding for an item or benefit or both. Precise information and data on C2C is both hard to get and hard to assess.

3.0 FRAUDULENT USERS IN E-COMMERCE APPLICATION

Trust is a particularly essential factor under states of vulnerability and hazard. The significance of trust is hoisted in web based business in light of the high level of vulnerability and hazard exhibit in most on line exchanges. In the present electronic universe of business, trust is the middle segment between the customer and the Internet Merchant. Analysts discovered trust imperative, particularly, in the connections amongst customers and e-merchants.

Fig 3.1 Relationship between Consumer and Internet Merchant

There is a solid connection between customer trust and security viewpoints that administer the entire exchange forms in a web based business site. As another type of business movement, internet business includes more vulnerability and dangers that customer trade since they are less notable to purchasers. Components that influencing trust in web based business for buyers incorporate security dangers, protection issue and absence of unwavering quality internet business forms when all is said in done. A buyer can't screen the wellbeing and security of sending touchy individual and money related data. Online business associations should look for innovative security component to shield itself from interruption and furthermore shield it client from being in a roundabout way attacked. There are two lines of protection for online business which are innovation arrangements and approach arrangements.

4.0 RELATED WORK

L. Azzopardi, M. Girolami, et.al [2003] an empirical study has been conducted investigating the relationship between the performance of an aspect based language model in terms of perplexity and the corresponding information retrieval performance obtained. It is observed, on the corpora considered, that the perplexity of the language model has a systematic relationship with the achievable precision recall performance though it is not statistically significant.

A.Ntoulas, M.Najork, et .al[2006] In this paper, to proceed with our examinations of "web spam": the infusion of misleadingly made pages into the web so as to impact the outcomes from web indexes, to direct people to specific pages for entertainment only or benefit. This paper thinks of some as beforehand UN portrayed methods for naturally recognizing spam pages, looks at the viability of these strategies in disconnection and when utilizing grouping calculations amassed. Whenever joined, our heuristics accurately recognize 2,037 (86.2%) of the 2,364 spam pages (13.8%) in our judged accumulation of 17,168 pages, while misidentifying 526 spam and non-spam pages (3.1%).

A. Klementiev, D. Roth, et.al.[2007] Numerous applications in data recovery, common dialect handling, information mining, and related fields require a positioning of examples regarding a predefined criteria rather than an arrangement. Moreover, for some such issues, numerous set up positioning models have been very much examined and it is alluring to consolidate their outcomes into a joint positioning, a formalism indicated as rank collection. This work exhibits a novel unsupervised learning calculation for rank collection (ULARA). Not with standing showing ULARA, we exhibit its adequacy on an information combination undertaking crosswise over impromptu recovery frameworks.

E. - P. Lim, V. - A. Nguyen, N.et.al. [2010] this paper intends to distinguish clients producing spam surveys or audit spammers. To distinguish a few trademark practices of survey spammers and model these practices in order to identify the spammers. To start with, spammers may target particular items or item bunches keeping in mind the end goal to amplify their effect. Second, they tend to veer off from alternate commentators in their evaluations of items. And propose scoring techniques to quantify the level of spam for every commentator and apply them on an Amazon audit dataset. Our outcomes demonstrate that our proposed positioning and directed techniques are compelling in finding spammers and outflank other pattern strategy in view of support votes alone. At last demonstrate that the distinguished spammers have more huge effect on appraisals contrasted and the unhelpful commentators.

5.0 PROBLEM FORMULATION

The existing framework just manages expectation and presumption graphs, here the diagrams will be in the typical arrangement to comprehend the information. In characterization, one is worried about relegating articles to classes on the premise of estimations made on these items. There are two primary viewpoints to arrangement: separation and bunching, or regulated and unsupervised learning. In unsupervised adapting (otherwise called group examination, class disclosure and unsupervised example acknowledgment), the classes are obscure an earlier and should be found from the information. Interestingly, in managed adapting (otherwise called segregate investigation, class expectation, and directed example acknowledgment), the classes are predefined and the assignment is to comprehend the reason for the order from an arrangement of marked articles (preparing or learning set). This data is then used to characterize future perceptions. The present article concentrates on the unsupervised issue, that is, on bunch investigation, however draws on thoughts from directed figuring out how to address the issue.

Problems in Existing Method

- ✓ Still now there are no methods in any social network to find out fake comments and ratings.

- ✓ Still More social networks are struggling to find fraud reviewers.
- ✓ Some classification methods like clusters and fuzzy are used in the existing method but both are not in efficient manner.
- ✓ The existing method can monitor the number of comments only, but it will not deals with any comments or rating.
- ✓ Still vendors are suffering due to competitor's fake comments.
- ✓ Many good products are rated poor due to business competitions.
- ✓ No higher end techniques are used.

6.0 THE PROPOSED SENEMB MODEL

The principle inspiration of this proposed framework is Sensitivity information examination. This is the proposed framework shows of how the vulnerability in the yield of a scientific model or framework (numerical or something else) can be allotted to various wellsprings of vulnerability in its contributions with REDDIT RANKING ALGORITHM. A related practice is vulnerability investigation, which has a more noteworthy concentrate on vulnerability measurement and engendering of vulnerability. In a perfect world, vulnerability and affectability investigation ought to be keep running couple.

The reddit positioning calculation emotionally supportive networks frequently observe information as information 3D shapes. The block is utilized to speak to information along some measure of intrigue. Despite the fact that called a "shape", it can be 2-dimensional, 3-dimensional, or higher-dimensional. Each measurement speaks to some quality in the database and the cells in the information shape speak to the measure of intrigue. For instance, they could contain a mean the quantity of times that quality blend happens in the database, or the base, greatest, aggregate or normal estimation of some trait. Inquiries are performed on the 3D square to recover choice help data.

Merits in Proposed System

- ✓ This is an entire new method for find out the fake reviewers in online market.
- ✓ It can be applicable for various social networks and various ecommerce applications.
- ✓ Advanced cluster methods are used in this application for clarity data retrieval.
- ✓ The proposed method will monitor all the user comment, number of comments per user and their commented text.
- ✓ Using the commented text, text categorization will be done using ANN.
- ✓ So fake and fraud reviewers will be no more in this application. All the comments from the fake reviewers will be terminated.
- ✓ Actual comments will be shown to the users.
- ✓ Users can buy any product by viewing proper reviews.

6.1 IMPLEMENTATION OF RANKING DATA ANALYSIS

Algorithm Implementation

Fig 6.2 RRA Algorithm

Step 1 : Counting total number of comments for a user during their post =TC;

Step 2: Calculating 1:8 Ration for Positive boost and 8:1 for negative boost;

Step 3: If 1<=8 & 1*8 = >8 means, the user may exceed with PC(Positive Comment) and 20 marks will be added in the fraudulent range, else No;

Step 4: If 8<=1 & 8*1=>1 means, the user may exceed with NC (Negative Comment) and 20 marks will be added in the fraudulent range, else No;

Step 5: Fraudulent Range will stored in an array namely FD[0]arr; check up to >60;

Step 6: If product purchased & PC >=5 FD[20]arr else 0, FD[0]arr;

Step 7: Also product purchased & NC>=5 FD[20]arr else 0,FD[0]arr

Step 8: check FD[num]arr >60

Step 9: Check not purchased PC >= 3 FD[20]arr if <3 FD[0]arr;

Step 10: As per previous check not purchased NC>=3 FD[20]arr if<3 FD[0]arr;

Step 11: check FD[num]arr >60 : These above given 6 are the important criteria as per opinion mining concept in the mining technique.

Step 12: Date of creation be DC: Initially DC will be in 0. Each day DC = i=i++. So that DC =[i+1];

Step 13: Go to step 3, Step 4, for Condition 1; FD[20]arr (r) FD[0]arr;

Step 14: Go to step 6, Step 7 for Condition 2; FD[20]arr (r) FD[0]arr;

Step 15: Go to Step 9, Step 10 for Condition 3; FD[20]arr (r) FD[0]arr;

Step 16: Considering Step 13, Step 14, Step 15 : if any of 2 conditions marked as FD[20]arr (r) FD[0]arr;

Step 17: In case of true FD[20]arr else FD[0]arr;

Step 18: Number of login = 0; if username and password is valid = Number Login = i+1;

Step 19 : Go to step 3, Step 4, for Condition 1; FD[20]arr (r) FD[0]arr;

Step 20: Go to step 6, Step 7 for Condition 2; FD[20]arr (r) FD[0]arr;

Step 21: Go to Step 9, Step 10 for Condition 3; FD[20]arr (r) FD[0]arr;

Step 22 : Go to 17; FD[20]arr else FD[0]arr;

Step 23 : In case of true FD[10]arr else FD[0]arr;

Step 24: Check number of same IP from DC >=3 : FD[10]arr else FD[0]arr;

Step 25: Calculating Fraudulent user:

Step 26: IF FD[>=60]arr Fraudulent user: TC, PC, NC,DC will be cleared from the DB.

Step 27: Else No change in DB.

DIRECTIVES

A directive is a special instruction on how asp.net should process the page. The most common directive is <%@ page %> which can specify many attributes used by the asp.net page parser and compiler.

REVIEW LENGTH (RL)

Review length decides the final and actual review of the user. Also review length is the average number of words present in a review. Usually the length of fake review will be on the lesser side because of the following reasons. Reviewer will not be having much knowledge about the product/business. Reviewer tries to achieve the objective with as few words as possible.

N-GRAM

An n-gram is a contiguous sequence of n items from a given sequence of text or speech. The items can be phonemes, syllables, letters, words or base pairs according to the application. These n-gram typically are collected from a text or speech corpus. In this project we use unigram and bigram as important features for detection of fake reviews. Unigram is an n-gram of size 1 and Bigram is an n-gram of size 2.

Unigram Frequency

Unigram frequency is a feature that deals with number of times each word unigram has occurred in a particular review.

Unigram Presence

Unigram presence is a feature that mainly finds out if a particular word unigram is present in a review.

Bigram Frequency

Bigram frequency is a feature that deals with number of times each word bigram has occurred in a particular review.

Bigram Presence

Bigram presence is a feature that mainly finds out if a particular word bigram is present in a review.

7.0 RESULTS AND DISCUSSION

FRAUDULENT USER DETECTION

In view of the past or past positioning records are initially actualize a basic and effective calculation which decides the main sessions of every Application. At that point, with the assistance of the examination of Applications positioning courses of action, when analyze false Applications and ordinary Applications, to recognize that there are distinctive positioning plan in each driving session. Hence, some misrepresentation confirmations from past positioning records of Applications, and 3 capacities are executed to get these positioning based extortion confirmations. In light of Application rating and audit history, we additionally create two sorts of

extortion confirmations, which mirror a few examples from Application authentic records. In Ranking Based Evidences, by assessing Application past/past positioning records, and distinguish which Application positioning practices in a headliner dependably contain a particular positioning example, which incorporates of 3 sorts of positioning stages, specifically, developing, saving and decay stage. In Rating Based Evidences, in addition, when Application has been issued, any clients who have downloaded the Application can give the evaluations for the predetermined issued Application. Also, the most imperative component of Application commercial is fundamentally in light of client rating.

Fig 7.1 Ranking fraud user

To show that in leading sessions ranking fraud is happened and given a procedure for mining main sessions for every Application from their past rating/ranking/review documents. For the detection of ranking fraud identify ranking-, rating- and review-based. Furthermore, a development based integration method has been proposed to combine all the evidences for calculating the quality of main sessions from Ecommerce Application. Here the distinct aspect is that by using the statistical hypothesis test can be used to model all of the evidences, thus it is easy to detect ranking fraud from domain knowledge with extended evidence. Lastly, prove the proposed system by performing test on real world Application where in Application data is collected from the Application store. Test on these parameters indicates the success of our work.

8.0 CONCLUSION AND FUTURE SCOPE

The system is similar to an advanced decision support system that provides useful transformation to the decision makers of an ecommerce review analyzers. Reddit ranking algorithms works perfect and produce accurate result as per expectation. Here after no need for people to read out all the reviews commented by the entire users, the system filters the anonymous and fake users from the database. In added with a clarity Graphical User Interface display is developed to view clarity information. Due to this information helps in making decisions regarding assignment of unusual transactions in the data system. The system required by the client based on their input in a faster manner. Since the Input given by the client is analyzed using the various data mining techniques and the result has been proved. Also the RRA proves the accuracy in finding a fraudulent user up to our expected level. To developing an E-commerce application using major functionalities like RRA's, GRM (General Ranking Method) are used. All the result has been verified the exact fraudulent user are suspected under the system.

Even the system is working well a according to the commitment, still need some enhancement to make the system more efficient and better. E commerce application has been used in this application, as the content management system, the system need to meet out the latest technology. As per the study the ecommerce application will work efficiently on cloud computing, our current system will be supporting on cloud computing. Future work will be based on the Green Computing. Green computing overcomes all the drawbacks in the cloud computing. Our system supports client server architecture also.

REFERENCES

[1] 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.

[2]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.

[3] A. Klementiev, D. Roth, and K. Small, "Unsupervised rank aggregation with distance-based models," in Proc. 25th Int. Conf. Mach. Learn., 2008, pp. 472–479.

[4] 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.

[5] 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.

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

[7] Z. Wu, J. Wu, J. Cao, and D. Tao, "HySAD: A semi-supervised hybrid shilling attack detector for trustworthy product recommendation," in Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2012, pp. 985–993.

[8] 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.

[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.

[15] 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.