International Journal of Engineering & Technology, 7 (4.19) (2018) 605-608

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

User Behaviour Analysis for the Merchandises Fairness Evaluation

Ghaidaa A. Al-Sultany[1], Saba Mohammed Hussain[2]

[1]*Department of Information Network, College of Information Technology, University of Babylon, Babil, Iraq*
[2]*Department of Information Network, College of Information Technology, University of Babylon, Babil, Iraq*
**Corresponding Author E-mail: ghaidaa.bilal@itnet.uobabylon.edu.iq*

Abstract

The development of digital technology leads to expansion of advertisements of goods and purchases online. The customer reviews and opinions affect in an essential way on the promotion of products and its reputation. In this paper, for the sake of distinguishing the user behavior in terms of his fairness and bias to the purchased merchandises, the user history has been analyze for extracting crucial features (Extreme rating, product's goodness, and user's past reviews). The extracted features help economic companies to exclude biased reviews for the fairness purposes of the evaluation of the product. In addition, they bring to light the identity of the user according to his past ratings and reviewing on purchased goods. The experiments in the research have shown encouraging results with respect to the values of the extracted features.

Keywords: *Extreme rating, user reviews, merchandises' goodness component; formatting; style; styling; key words.*

1. Introduction

Today, digital technology becomes an urgent necessity for online shopping. Therefore, during the last decades, the Internet played an important role in the success of those goods [1].
Many companies announce their goods through the Internet such as eBay, Amazon, AliExpress….etc. They pass their purchased merchandises to their customers to review them. Reviews can affect the people opinions on either positive or negative way. Hence, reviews on products have become a significant source of information for deciding the sale and buying goods.
Some peoples are biased in their rating and reviewing the merchandises [2]. Such this kind of people cannot reflect the real values of the purchased goods, therefore, most economic companies tried to exclude such this type of biased reviews for the fairness purposes on the evaluation of the product.
The quality of the product depends on the user evaluated. Therefore, there are many systems determined between fairness users in order to maintain the popularity of products from untruthful user effects. It depends on various measures in classifying process between users or reviewers. The profile of users gives full scope about the activity of users to help systems to determine the fairness. In addition, the profile of products gives purchases enough information about goods. On the other hand, the profile of the reviewers and products help the purchases to select good goods when buying and selling [3]. This necessity motives us to the proposed new system for extracting a profile of reviewer and product from analysis reviews and reviewers on the products. For the sake of, support the system with a complete observation of users' activity. In addition, support systems increase the accuracy of the system. A good analysis of user history increases system confidence in classification.
The remainder of the paper is organized as follows: In Section 2 we explain the related work. Section 3 presents the preprocessing for the dataset. Section 4 explains the methodology of the work for four indicators.in sections 5 descriptions for the dataset and section 6 discusses results. Finally, Conclusion and future works stated in Section 7.

2. Related Work

The academics submit many spam detection techniques such as [4,5and6]. The main task of using these techniques is recognized between fake and authentic reviews by extracting some features from the history of authors. Manisha et al., in [7], the authors suggested a model save a history of reviewers to identify the fake and trust reviewers. The system makes a decision based on some features extracted from the user and product, such as ID, user IP Address, brand ID, product ID, product rate, and review. When the user posts a new review for a product, the system checks whether a user previously saved in the model (i.e., approximately has the same above features) blocked the user or sent an error message to him .
Srijan at. el. in [8] proposed a new algorithm (Rev2) depend on measuring the fairness of a user, reliability of a rate, and goodness of a product for sake of classifying reviews into fake and trust. Kyungmin et. al. in [9] identified two types of review patterns, authentic and fake based on configurations among reviewers and review content elements. Each pattern has unique characteristics. The authentic reviews' patterns explain by personality theory (i.e., five factors model). The fake review pattern explains by HSM and ELM. As well as, they explain what fake review is with IMT as theoretic background. The results, which are based on the fsQCA method, identify the combinations of configurations for authentic and fake reviews.
The study by Jitendra and Smriti in 2017in [10], this work was ordered based on techniques, reviewers' features, datasets of

products and reviews used in reviewers' spammers' detection. Furthermore, most effective feature sets were assemble for building the model. Thus, sentiment analysis was incorporate in the detection process. In order to obtain the best performance, some well-known classifiers were apply on a labeled dataset. For unlabeled data, clustering used after desired attributes were computed for spam detection.

Atefeh et. al. in 2016 in[11], proposed a review spam detection approach. They employed authors' activeness and rating behaviors as well as context similarity of reviews in each captured interval in order to assign spam scores to the reviews and distinguish fake reviews from real opinions.

One of the important features that every system depends on classifying users into spammer and non-spammer on the activity of these users. It could extract from the user profile. Therefore, the good analysis of user profile increases the accuracy of the system in its classification process.

Every proposed system suffers from certain drawbacks preventing it to identify all the harmful spam reviews such as of these challenges, collecting a huge amount of unlabeled data (raw reviews) is a feasible process. However, the acquisition of labeled data is difficult and costly. In addition, the quality of training data plays a critical role in developing accurate decision-making. Moreover, many of these approaches require a ready set of reviews to detect profile user. The literature shows that the others are not as reliable as to be confidently used in real situations.

3. Preprocessing Dataset

Düzenlemeye Data preprocessing is an important step in the data mining process [12]. Data mining and machine learning techniques, primarily those for web and text mining, offer an exciting contribution to detecting fraudulent reviews. We can define web mining is "the process for finding useful information and relations from the contents available on the web by largely relying on the available machine learning techniques and methods". Big data maybe need to be clearbefore used for data mining. Purifying or cleaning process analyses [13] missing values, different formats of date/time, and empty or garbage values. The proposed system analyses dataset which product by Amazon [16]. The original format of data is JSON (JavaScript Object Notation). The system converts the string of JSON data to a Python data structure for enable to save data into CSV format. Each attribute post in an individual column such as (User ID, User Name, Date, Product Name, Rate, Review) in new CSV file. The date of each review converts into standard date format (*month/ day/year)* for sake of calculating the life of review and number of reviews per time.

4. Methodology

The proposed system express profile of the user based on analysis history of the user. In work, extract a number of features related with user or reviewer such as fairness, Number of review per product, and the content similarity for all user's reviews. These attributes help a system to determine the user identity. In extracting user profile, the system depends on analysis data of the user (reviews, rate, and a number of the rate per product). Figure (1) shown the proposed system:

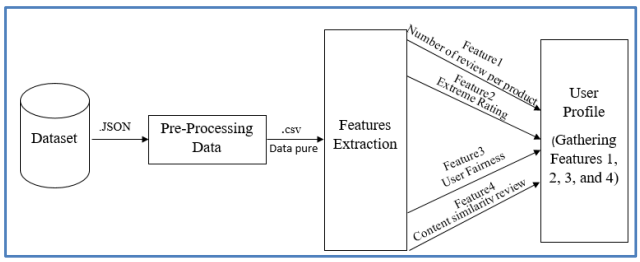

Fig. 1: Proposed system

4.1. Extreme Rating:

Users may give extreme ratings to his purchases (e.g. 1 or 5 stars) for the purposes of increasing or decreasing the mean score of a product. In this research define extreme rating has been computed with respect to the user purchases history to analyze his behavior to the products in total. The user extreme rating has been calculated using the absolute difference between the ratio of user positive rates (5, 4, and 3) and the ratio of negative rates (2, and 1) from his total rates as stated in equation 1,2 and 3 respectively.Function (1) calculates the positive fairness of user (P_i^+)

$$P_i^+ = \frac{no.of\ good\ rating}{Total\ rate} \quad (1)$$

Function (2) calculates the negative fairness of user (P_i^-)

$$P_i^- = \frac{no.of\ bad\ Rate}{Total\ rate} \quad (2)$$

Function (3) calculates the extreme rating (Exr_i)

$$Exr_i = |P_i^+ - P_i^-| \quad (3)$$

Figure (2) illustrates the pseudo code of extreme rating

1: Data ← User Purchased Merchandises History
2: Users ← All user
3: While $U_i \in$ Users do
4: $Positive(U_i) \leftarrow$ Data $(all\ rates\ of\ U_i) \geq 3$
5: $Negative(U_i) \leftarrow$ Data $(all\ rates\ of\ U_i < 3$
6: $extreme\ Rate(Ui) \leftarrow \left|\frac{Positive(U_i)-Negative(U_i)}{total\ Rate\ of\ U_i}\right|$
7: return $extreme\ Rate$

Fig. 2: Pseudocode of extreme rating

4.2. Content Similarity

Some users often replicate the same review, which could detect by considering the overall content similarity of their reviews. The users content similarity assigned to them based on their reviews. In addition, the users have many reviews that are more likely to have a higher proportion of content similarity score [14]. To check the similarity between users' reviews, the content similarity has been calculated through using many of different ways [12] such as Cosine, Jaccard, and Term Frequent. Inverse Document Frequent (TF.IDF (. In this paper, TF.IDF was implemented as one of the most popular methods that were concentrated on evaluating how important the words' significance in documents as illustrated in equations (4, 5and6) in [13]. TF.IDF is a very interesting way to convert the textual representation of information in order to sparse features to find the similarity between the texts, because it is more accurate as shown in Figure (3). Function (4) calculates the TF, where n document and f_{ij} to be the frequency, suppose term i appears in n_i of the N documents in the collection:

$$tfi,j = \frac{\mathrm{ni,j}}{\sum_k \mathrm{ni,j}} \quad (4)$$

Function (5) calculates the IDF

$$idf(w) = \log\left(\frac{N}{dfi}\right) \quad (5)$$

Function (6) calculates the w

$$wt,d = TFt,d \log\left(\frac{N}{DFt}\right) \quad (6)$$

Where N documents. Define f_{ij} to be the frequency (number of occurrences) of term (word) the $i$the n document$j$. Then, define the

term frequency TF_{ij} to be suppose term i appears in n_i of the N documents in the collection. Figure (3) showing the structure of TFIDF

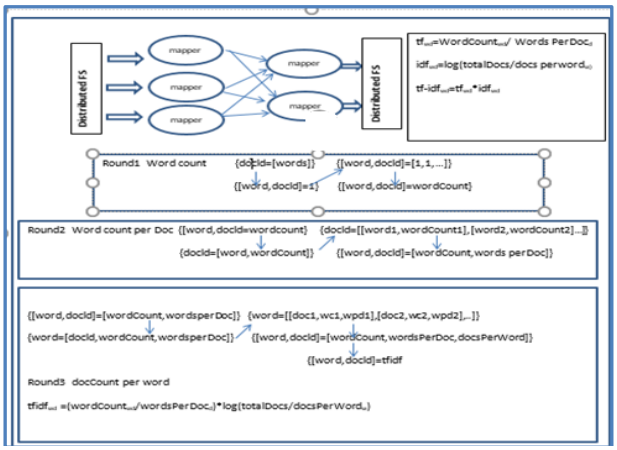

Fig. 3: Structure of TFIDF

4.3. Number of Reviews per Product (NRP)

Due to the impact of excessive reviewing, we also consider a reviewer's relevant behavior in past reviewed products. According to [15] measure the average number of reviews a reviewer a(ri) have written per product by dividing the size of his reviewing history Hista,j over the number of reviewed products $n_{a,p}$. Function (7) calculates the NRP.

$$NRP(a(ri)) = \frac{Hist\ a,j}{na,p} \quad (7)$$

Figure (4) shown the pseudo code of the number of Reviews per Product (NRP).

1: Data ← User Purchased Merchandises History
2: Users ← All user
3: Products ←All Products
3: While Pri ∈ Products do
4: $Hist\ (\text{Users}_i, \text{Products}_j) \leftarrow Data(\text{Users}_i, \text{Products}_j)$
5: $Nap_{Pri} \leftarrow Data(\text{Products}_j)$
6: $NRP(\text{Users}_i, \text{Products}_j) \leftarrow \frac{Hist\ (\text{Users}_i, \text{Products}_j)}{Nap_{Pri}}$
7: return NRP

Fig. 4: Pseudo code of Number of Reviews per Product (NRP)

4.4. Goodness of products

A goodness score is a unique number representing the most expected rating from an authentic user that given for the product. Naturally, a good product would get a high number of positive ratings, and a bad product would receive high negative (G(p) =0), and one to bad products (G(p) =1). the systems select one comment per user on the product. The work in this paper gives zero to good products. Function (13) calculates the goodness of products.

$$\Delta g = T_g - U_g \quad (8)$$

where, T_g are total good rates, U_g is a unique user rate

$$\Delta b = T_b - U_b \quad (9)$$

Where, T_b are total bad rates, U_b is a unique user rate

$$D = \Delta g - \Delta b \quad (10)$$

$$g = (U_g + 1)/((T_r - D) + 1) \quad (11)$$

$$b = (U_b + 1)/((T_r - D) + 1) \quad (12)$$

$$G(P) = \begin{cases} 0 & g > b \\ 1\ otherwise \end{cases} \quad (13)$$

Where T_r is all rating that is give from users to prodthe uct (good and bad with frequent). The result from above if $G>B$ then put 0 else put 1to Goodness.Figure (5) shown the pseudo code of Goodness.

1: $Data \leftarrow User\ Purchased\ Merchandises\ History$
2: $Users \leftarrow All\ user$
3: $Products \leftarrow All\ Products$
4: While $Pr_i \in$ Products do
5: $user_i \in Users$
6: $Rate_{pr_i} \leftarrow Data(Pr_i, user_i)$
7: $GoodRate_{Pri} \leftarrow Rate_{pr_i} \geq 3$ // pickup one rate of each user for a product Pr_i
8: $BadRate_{Pri} \leftarrow Rate_{pr_i} < 3$ // pickup one rate of each user for a product Pr_i
9: $delta_{Good} \leftarrow No.total\ GoodRate_{Pri} - unique\ No.GoodRate_{Pri}$
10: $delta_{Bad} \leftarrow No.total\ BadRate_{Pri} - unique\ No.BadRate_{Pri}$
11: $Good_{pri} \leftarrow \frac{GoodRate_{Pri}+1}{total\ Rate_{Pr_i} - (delta_{Good}+delta_{Bad})+1}$
12: $Bad_{pri} \leftarrow \frac{BadRate_{Pri}+1}{total\ Rate_{Pr_i} - (delta_{Good}+delta_{Bad})+1}$
13: if $Good_{pri} \geq Bad_{pri}$ then $Goodness_{pri} \leftarrow 0$
14: else $Goodness_{pri} \leftarrow 1$
15: return $Goodnes$

Fig. 5: Pseudo code of Goodness

5. Dataset Description

We analysis that one of Amazon dataset [16] described as follows:

- Number of samples: 5000 records
- Number of users: 688 users
- Number of products: 1279
- productsDataset gathering date: from 2007 – 2014
- Rate interval : [1, 5]

6. The Results Discussion

The analysis of the dataset according to propose a system to know the behavior of users from them profile has been done in this stage. The first indicator is user content similarity, the ratio of content review similarity is divided between (0 and 1), in order to show the ratio of user density. Where the ratio is close to 0 is the best proportion and the similarity ratio which closer to 1 is worst. Figure (6) shown the content review similarity of users.

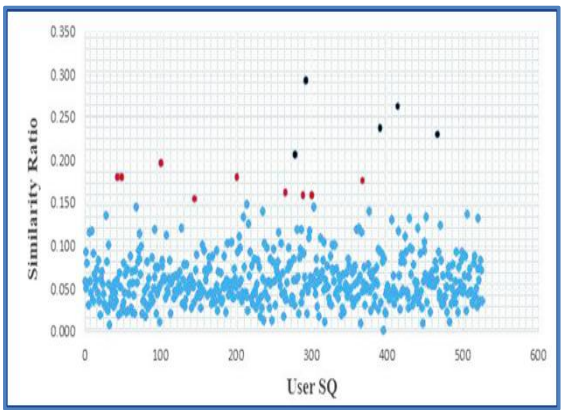

Fig. 6: Density of user content similarity

The result of second indicator Fairness, in this, take all the user to have a rating on the products more than (10 rates). The ratio that nearest to 0 is the best proportion which gives indicate the good fairness of user and value of fairness which closer to 1 is the worst ratio that gives a negative indicator about the user. Figure (7) illustrate the fairness of users.

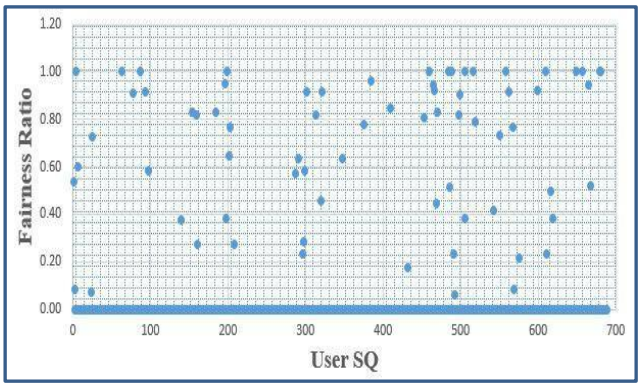

Fig. 7: the Fairness of user

The third indicator is the goodness of product, in this case, we take all the product in the dataset (1279), where to give 94% are good products and 6% are bad products according to evaluate of users .as shown in figure (8) :

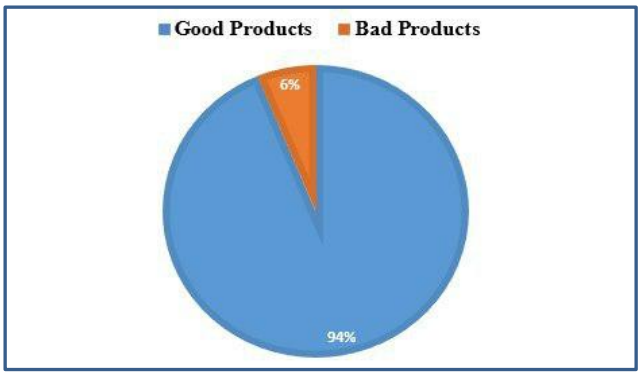

Fig. 8: Goodness of products

Figure 9 illustrates the behavior of four users according to the features that extract from users' data(no. of review per product, content similarity, and fairness), where users (a, and b) are normal and (c, and d) are suspicious.

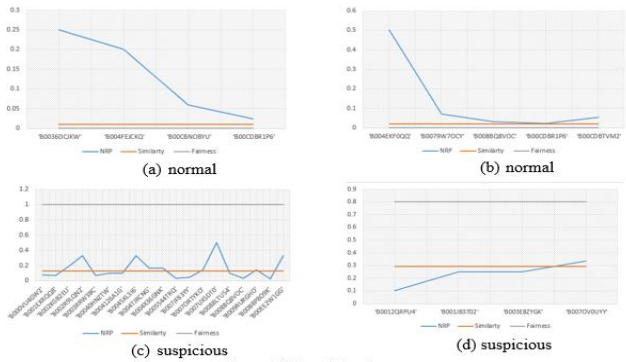

Fig. 9: User behavior

7. Conclusion

The recommendation systems, which classify users into fake and truth, depend on user behavior. The strong system which could analysis user profile well. The attributes that extract from the user's data sport the classification system in the mission entrusted to it. The spammers take deferent styles for sake of hiding itself from the discovery. Therefore, the variant features that extract from user's data help detected spammers and minimize the distraction. The results of this research are considered as input to many research and applications, especially social networking sites, which are good results compared to previous research, which relied on less than these characteristics to determine the history of the user. In the future, it is possible to combine more features to reveal the identity of the user more efficiently.

References

[1] Hao Tian and Peifeng Liang, " Improved Recommendations Based on Trust Relationships in Social Networks", Future Internet 2017
[2] Michael Crawford, Taghi M. Khoshgoftaar, Joseph D. Prusa, Aaron N. Richter, and Hamzah Al Najada, " Survey of review spam detection using machine learning techniques", Crawford et al. Journal of Big Data (2015) 2:23.
[3] Eka Dyar Wahyuni and Arif Djunaidy, " Fake Review Detection From A Product Review Using Modified Method OF Iterative Computation Framework", Web of Conferences, DOI: 10.1051/ MATEC 58 03003 (2016).
[4] Jindal, Nitin, and Bing Liu. "Opinion spam and analysis." Proceedings of the 2008 international conference on web search and data mining. ACM, 2008.
[5] Beel, Joeran, and Bela Gipp. "Academic search engine spam and Google Scholar's resilience against it." Journal of electronic publishing 13.3 (2010).
[6] Carpinter, James, and Ray Hunt. "Tightening the net: A review of current and next-generation spam filtering tools." Computers & Security 25.8 (2006): 566-578.
[7] Manisha Singh, Lokesh Kumar, and Sapna Sinha, " Model for Detecting Fake or Spam Reviews", (Springer Nature Singapore Pte Ltd. 2018).
[8] Srijan Kumar, Bryan Hooi, Disha Makhija, " REV2: Fraudulent User Prediction in Rating Platforms", WSDM 2018, February 5–9, 2018, Marina Del Rey, CA,
[9] Kyungmin Lee1, Juyeon Ham2, Sung-Byung Yang3, and Chulmo Koo, " Can You Identify Fake or Authentic Reviews? A fsQCA Approach", © Springer International Publishing AG 2018
[10] Rupesh Kumar Dewang and Anil Kumar Singh, " State-of-art approaches for review spammer detection: a survey", J Intell Inf Syst (2018).
[11] Atefeh Heydari, Mohammad Ali Tavakoli a, Naomie Salim and Zahra Heydari, " Detection of review spam: A survey", Expert Systems with Applications 42 (2015) 3634–3642.
[12] Han, Jiawei, Jian Pei, and Micheline Kamber. Data mining: concepts and techniques. Elsevier, 2011.
[13] Uma, K., and M. Hanumanthappa. "Data Collection Methods and Data Pre-processing Techniques for Healthcare Data Using Data Mining.", International Journal of Scientific & Engineering Research Volume 8, Issue 6, June-2017,
[14] Atefeh Heydari, Mohammad Ali Tavakoli and Naomie Salim., " Detection of fake opinions using time series", Expert Systems With Applications 58 (2016) 83–92.
[15] Ioannis Dematis (&), Eirini Karapistoli, and Athena Vakali, " Fake Review Detection via Exploitation of Spam Indicators and Reviewer Behavior Characteristics", © Springer International Publishing AG 2018.
[16] http://jmcauley.ucsd.edu/data/amazon , 2018