Comparative Analysis of Online Rating Systems

Mohammad Azzeh
Faculty of Information Technology
Applied Science Private University
Amman, Jordan POBOX 166

***Abstract*—Online rating systems serve as decision support tool for choosing the right transactions on the internet. Consumers usually rely on others' experiences when do transaction on the internet, therefore their feedbacks are helpful in succeeding such transactions. One important form of such feedbacks is the product ratings. Most online rating systems have been proposed either by researchers or industry. But there is much debate about their accuracies and stability. This paper looks at the accuracy and stability of set of common online rating systems over dense and sparse datasets. To accomplish that we used three evaluation measures namely, Mean Absolute Errors (MAE), Mean Balanced Relative Error (MBRE) and Mean Inverse Balanced Relative Error (MIBRE), in addition to Borda count to assess the stability of ranking among various rating systems. The results showed that both median and Dirichlet are the most accurate models for both sparse and dense datasets, whereas the BetaDR model is the most stable model across different evaluation measures. Therefore we recommend using Dirichlet or BetaDR for the products with few number of ratings and using the median model with products of large number of ratings.**

Keywords—Online rating systems; reputation models; comparative analysis; decision making; e-commerce

I. Introduction

Online rating systems play a vital role in most ecommerce applications. They help users to facilitate their decisions while they perform internet transactions [1], [4]. The online rating system is responsible for collecting, processing and aggregating ratings given for a specific product. The main challenge that faces the online rating systems is how to aggregate the collected ratings for a specific product in way that can reflect its real quality [13]. In practice, most of the well-known ecommerce portals such as eBay, Amazon, etc. use their own methods to compute the quality of product. But some other portals use the simplest aggregation method which is the Naïve average methods (i.e. mean, median and mode). In contrast, many authors proposed different method to compute product score based on statistical and machine learning methods. The accuracy of such methods depends mainly on the user satisfaction about the results achieved [14]. This satisfaction is difficult to be measured because most ecommerce application don't provide a tool to evaluate the user satisfaction, and whether the given aggregate rating help them in performing the successful transaction. The rating aggregation methods in literature can be divided into four groups, Naïve models, weighted average models, Fuzzy models and probabilistic models. The weighted average models are the widely used among researchers, where the weights are derived from historical user data or time factor. These weight values work as discount factors to reflect different aspects of users' behavior such as their reliability, trustworthiness and credibility in providing rating. One of the common problem that faces rating systems is unfair ratings that biases aggregate scores for some products.

This paper attempts to look at the accuracy and stability of the most common online rating systems over dense and sparse datasets. Practically, not all methods perform well over dense or sparse datasets. This fact has been confirmed by almost all previous studies because each model attempts to treat a specific limitation in previous rating systems. To best of our knowledge, there is no systematic procedure has been conducted to compare and evaluate different online rating systems in terms of accuracy and stability. The proposed research questions are:

RQ1: Is there any one method that can perform stably well under all conditions?

RQ2: Which group of methods is more appropriate for dense datasets?

RQ3: Which group of methods is more appropriate for sparse datasets?

This paper is structured as follows: section 2 presents the literature and overview of existing online rating systems. Section 3 introduces the experimental methodology and comparison procedure. Section 4 presents the obtained results, finally we end up with the conclusions in section 5.

II. Overview

Online rating system receives ratings from users as input to compute the aggregate score of product. Given a set of users $U = \{u_1, u_2, u_3, \ldots, u_n\}$ where each user rated at least one product, also given a set of products $P = \{p_1, p_2, p_3, \ldots, p_m\}$ where each product received at least one rating, the intersection between user u_i and the product p_j is the rating r_{ij} such that $1 < r_{ij} \leq k$. k is the maximum rating level for rating system. $\overline{p_j}$ is the ratings average of product p_j, and $\bar{P}$ is the average of all ratings in the dataset. Indeed, Naïve methods such as arithmetic mean (see Equation 1) and median are the most common used methods. Garcin et al. [15] compared between Naïve methods and other rating systems. They revealed that the median is the most accurate method. In contrast, other studies [8], [9] showed that the naïve methods are ineffective because they are easily influenced by unfair and malicious ratings and cannot discover trend emerging from recent ratings.

IMDb is another famous online rating system that uses true Bayesian estimation to calculate the aggregate product score as shown in (2). The exact implementation of this model is still unpublished in order to keep the policy effective.

$$\bar{p_j} = \frac{1}{n}\sum_{i=1}^{n} r_{ij} \tag{1}$$

$$IMDb\_score_j = \frac{n}{(n + MinR)} \times \bar{p_j} + \frac{MinR}{(n + MinR)} \times \bar{P} \tag{2}$$

Where n is the number of ratings received for product p_j. *MinR* is the minimum number of rating count required to appear on the top 250. IMDb usually uses *MinR*=2500.

In literature, the weighted average models are the widely used models, where the weights are computed based on either time or user data. Josang and Haller [5] introduced the age of rating as discount factor in computing and aggregating rating, where old ratings receive less weight than recent ratings because they are not informative. The main problem with this model is which time unite (i.e. day, week, month, year) should be considered with this function. Another time discount function used is the number of past transactions instead of using the ratings age [10]. Leberknight et al. [8] stated that the naïve methods are good when there is clear trend of ratings over time, but when the ratings do not have that trend one should involve the volatility of ratings as discount factor to compute the product score. They proposed discount function based on rating volatility, but they ignored the importance of other factors such as trustworthiness and credibility of users. On the other hand, many online rating systems use users' data to measure their reliability, credibility and trustworthiness and reflect that as weight during aggregation process [12]. In this direction, Riggs et al. [11] defined the reliability of a user by his ability to provide rating that is very close to the current ratings average. They defined a measure to calculate that closeness and use it with their weight average model. Lauw et al. [7] studied the leniency of user while they rate products. They proposed a function that can calculate the leniency and strictness of user and reflect that as weight. They classified users into two classes (lenient and strict) based on leniency variable l_i as shown in Equation 3, such that if $l_i < 0$ then reviewer is strict, otherwise reviewer is lenient. This model is called LQ.

$$l_i = \frac{1}{|u_i|}\sum_{j=1}^{|u_i|}\left(\frac{r_{ij} - q_j}{r_{ij}}\right) \tag{3}$$

$$q_j = \frac{1}{|e_j|}\sum_{i=1}^{|e_j|} r_{ij} \times (1 - \alpha \cdot l_i) \tag{4}$$

Where q_j is the initial quality of the item *j* which is usually the average of ratings. l_i is the leniency of the reviewer. $\alpha \in [0, 1]$ is a compensation factor determined by expert. Abdel-Hafez et al. [1], [16] used Beta distribution function to compute ratings weights. Their model is called BetaDR. The product ratings should be first sorted from smallest to largest and scaled as shown in (5). The beta distribution function has advantage such that it can change its shape based on the rating distribution. Therefore they controlled the shape of the function by two variables α and β as shown in (6). Finally, the product score is measured as shown in (7).

$$x_i = \frac{0.98 \times i}{n - 1} + 0.01 \tag{5}$$

$$Beta(x_i) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x_i^{\alpha - 1}(1 - x_i)^{\beta - 1} \tag{6}$$

$$BetaDR = \sum_{l=1}^{k}(L \times w_l) \tag{7}$$

Where Γ is the gamma function, and α and β are Beta distribution parameters that are determined based on mean and distribution of ratings. L is rating level (i.e. 1, 2, … *k*). w_l is the summation of normalized Beta weight for the target level. Jøsang et al. [6] introduced a reputation model based on Dirichlet probability distribution as shown in Equations 8 and 9. This model is a generalized form to their previous model and takes the rating counts in calculation. The model works well with good accuracy over sparse datasets because it involves factors that can treat uncertainty in the data.

$$\overrightarrow{S_y}: \left(S_y(i) = \frac{R_y(i) + Ca(i)}{C + \sum_{j=1}^{k} R_y(j)} ; | i = 1 \dots k \right) \tag{8}$$

$$score = \sum_{i=1}^{k} v(i)S(i);\ \ where\!: v(i) = \frac{i - 1}{k - 1} \tag{9}$$

where $\overrightarrow{S_y}$ represents the score vector of each rating level, $S_y(i)$ represents the probability that one agent gives rating *i* to agent *y*. C is a constant value, and (i) is the base rate, which equals to 1/*k*. $R_y(i)$ is the number of ratings of the level *i*.

Bharadwaj et al. [2] used the ordered weighted averaging method with fuzzy computation as part of their trust model to aggregate rating as shown in Equations 10 to 12. According to them, the reputation of a reviewer is defined as the accuracy of his prediction to other reviewer's ratings towards different items. Recently, Liu et al. [9] proposed several factors to identify unfair ratings. These factors are combined together using Fuzzy Logic System based on human predefined rules. The output of Fuzzy logic system is the discount weight of rating.

$$score_j = \sum_{i=1}^{n} W_i \times r_{ij} \tag{10}$$

$$W_i = Q\left(\frac{i}{n}\right) - Q\left(\frac{i - 1}{n}\right) \tag{11}$$

$$Q(r) = \begin{cases} 0 & 0 \le r \le 0.3 \\ 2 \times r - 0.6 & 0.3 < r \le 0.8 \\ 1 & 0.8 < r \le 1 \end{cases} \tag{12}$$

III. Experimental Setup

A. Datasets

Most authors used public datasets to validate their models which allow them to generalize the extracted knowledge. In this paper we continue that approach to facilitate the replication studies in future. Two stable versions of MoviLens datasets have been used namely, 100K and 1M [3]. Both datasets have large number of ratings which are considered dense datasets as shown n Table1. To compare online rating systems over sparse datasets, we extracted new three datasets from the original 1M dataset, where each new dataset contains randomly selected 4, 6 and 8 user ratings respectively. These datasets are called 1M4, 1M6 and 1M8. The characteristics of all datasets are shown in Table 1.

TABLE I. Datasets Characteristics

Dataset	#User	#Movies	#ratings
100K	943	1682	100,000
1M	6040	3706	1,000,209
1M4	6040	920	24,160
1M6	6040	1286	36,240
1M8	6040	1625	48,320

B. Evaluation measures

Evaluation measures are used to assess the accuracy and stability of online rating systems. To measure the accuracy of a model we used three measures, Mean Absolute Errors (MAE), Mean Balanced Relative Error (MBRE) and Mean Inverse Balanced Relative Error (MIBRE). These measures have been selected as they are not biased. The MAE assesses, for each product, the closeness of the generated score to the actual ratings for a product as shown in Equation 13. Both MBRE and MIBRE compute the relative accuracy of the generated scores as shown in Equations 14 and 15.

$$MAE = \frac{1}{m}\sum_{j=1}^{m}\frac{\sum_{i=1}^{n}|r_{ij} - score_j|}{n} \tag{13}$$

$$MBRE = \frac{1}{m}\sum_{j=1}^{m}\left(\frac{1}{n}\sum\nolimits_{i=1}^{n}\frac{|r_{ij} - score_j|}{\min(r_{ij}\ ,\ score_j)}\right) \tag{14}$$

$$MIBRE = \frac{1}{m}\sum_{j=1}^{m}\left(\frac{1}{n}\sum\nolimits_{i=1}^{n}\frac{|r_{ij} - score_j|}{\max(r_{ij}\ ,\ score_j)}\right) \tag{15}$$

Where $score_j$ is the aggregated score for product p_j. m is the number of products in the testing data.

C. Experimental procedure

As mentioned in the literature, there are many models have been proposed to aggregate online ratings. In this study we used eight state-of-art models are: Mean, Median, BetaDR [1], Bayesian [6], Dirichlet [5], IMDb, Fuzzy rating [2] and LQ [7]. For comparison purpose we used 10-Fold cross validation. This procedure divides the dataset into 10 groups of training and testing data. Each group has 90% of the data as training data and 10% as testing data. The training data is used to build the online rating system, while the testing data is used to evaluate the model. The validation is running 10 times, one time for each group. In each run we record the MAE, MBRE and MIBRE for test ratings. The fundamental idea of using this validation technique is that a reputation score that is produced from training dataset is considered accurate if it is very close to actual ratings in the testing dataset. To measure the stability for each model across different evaluation measures, we rank all models according to their accuracy in terms of MAE, MBRE and MIBRE over all datasets. Then we run Borda count method over all datasets, dense datasets and sparse datasets respectively. Borda count is voting ranked method used to rank various candidates based on the ranks provided by voters. This method is simple and very common in decision making area. First we evaluate the stability of all models over all datasets across all evaluation measures. Then in the second round we evaluate the stability over only dense datasets, then finally over sparse datasets. In all cases the evaluation measure work as voters.

IV. Results and Discussion

This section presents the results of comparisons among different online rating systems. Table 2 shows the MAE results over all datasets. From the results we can notice that the differences between all models are nearly negligible, except for LQ model where it is extreme over both dense and sparse datasets. It is interesting to know that Naïve models produce accurate results in comparison to more sophisticated models such as Bayesian and LQ. For the dense datasets (i.e. 100K and 1M) the median model produces the more accurate results, while for sparse datasets the Dirichlet and BetaDR are more accurate. This results confirmed previous findings that confirm that both Dirichlet and BetaDR were originally proposed to handle sparse datasets that contain very few ratings. In spite of that, the median model still produces comparable accuracy to Dirichlet model over all sparse datasets.

TABLE II. Mean Absolute Error Results

Dataset	Mean	Median	BetaDR	Bayesian	Dirichlet	IMDb	Fuzzy	LQ
100K	0.905	0.886	0.892	0.902	0.892	0.906	0.919	1.021
1M	0.841	0.810	0.832	0.844	0.841	0.855	0.848	0.962
1M4	0.877	0.876	0.872	0.882	0.883	0.909	0.887	0.982
1M6	0.911	0.907	0.906	0.926	0.886	0.908	0.916	1.023
1M8	0.907	0.897	0.901	0.902	0.883	0.909	0.921	1.007

To perform further investigations, we run the analysis using MBRE and MIBRE evaluation measures. Table 3 shows the results of MBRE over all datasets. Similar to Table 2, the accuracy results are close. Generally, we can observe that the Dirichlet model is the most accurate model over both dense and sparse datasets. Table 4 suggests that the median model is the most accurate model over all datasets. This variation in the results confirm that both median and Dirichlet models are the most accurate models for both sparse and dense datasets. Based on above analysis we can recommend using the median model because it has simple implementation than Dirichlet and can produce comparable to Dirichlet and better than many sophisticated models.

TABLE III. MBRE RESULTS

Dataset	Mean	Median	BetaDR	Bayesian	Dirichlet	IMDb	Fuzzy	LQ
100K	0.477	0.491	0.476	0.480	0.464	0.478	0.495	0.549
1M	0.418	0.422	0.419	0.429	0.416	0.430	0.431	0.463
1M4	0.409	0.425	0.414	0.418	0.395	0.416	0.426	0.457
1M6	0.428	0.439	0.428	0.445	0.395	0.410	0.434	0.487
1M8	0.421	0.430	0.425	0.425	0.394	0.411	0.438	0.506

TABLE IV. MIBRE RESULTS

Dataset	Mean	Median	BetaDR	Bayesian	Dirichlet	IMDb	Fuzzy	LQ
100K	0.251	0.240	0.246	0.248	0.248	0.249	0.250	0.288
1M	0.233	0.218	0.229	0.231	0.233	0.235	0.231	0.268
1M4	0.221	0.216	0.218	0.218	0.225	0.228	0.220	0.250
1M6	0.228	0.223	0.225	0.229	0.226	0.228	0.226	0.261
1M8	0.228	0.220	0.225	0.225	0.224	0.229	0.228	0.258

To analyze the stability of all models over all datasets and both sparse and dense datasets, we first rank all models over each dataset individually and over each evaluation measure. Then we apply the Borda count method. Table 5 presents the ranking stability of all models over dense and sparse datasets. From the results of ranking we can notice that the BetaDR is the most stable model over all datasets and especially over dense datasets across different evaluation measures, whereas the Dirichlet model is the most accurate model over sparse datasets. Generally, we can notice that the top three models in the table (i.e. BetaDR, Dirichlet and median) are the most stable models. The results obtained surprisingly suggest that the BetaDR is better than both Dirichlet and median over all datasets. In contrast, we can observe that the sophisticated models such as Fuzzy and LQ are not accurate as they occupy the last position over all datasets and across all evaluation measures. Also the commonly used mean model occupies mid positions with unstable ranking across all evaluation measures.

TABLE V. RANKING STABILITY

Rank	All datasets	Dense datasets	Sparse Datasets
1	BetaDR	BetaDR	Dirichlet
2	Dirichlet	median	BetaDR
3	median	Dirichlet	median
4	mean	Bayesian	mean
5	Bayesian	mean	IMDB
6	IMDB	IMDB	Bayesian
7	Fuzzy	Fuzzy	Fuzzy
8	LQ	LQ	LQ

Finally we revisit the proposed research questions in this study:

RQ1: Is there any one method that can perform stably well under all conditions?

Ans. Actually, there is no accurate answer because the difference among all models are negligible, but we can say that median and Dirichlet models produce the most accurate results as shown in Tables 2, 3 and 4.

RQ2: Which group of methods is more appropriate for dense datasets?

Ans. From Table 5 we can see that both BetaDR and median models are the most stable and accurate models over dense datasets.

RQ3: Which group of methods is more appropriate for sparse datasets?

Ans. Similar to previous answer, we can observe that Dirichlet and BetaDR are the most accurate and stable models over sparse datasets. This is not surprising results because the purpose of construction of both models was to treat the sparse datasets. Also both models are good for new rating system that has few numbers of ratings.

V. CONCLUSIONS

Online rating system is a helpful tool to facilitate user decision in conducting online transactions. However, the accurate rating system can let user choose the correct product which leads to better user satisfaction. Many models have been proposed in literature, but their accuracy are subject to the degree of helpfulness. In this paper we conducted a comparative analysis for the widely used online rating systems to investigate their accuracies and stability over dense and sparse datasets. Three evaluation measures in addition to Borda count method have been used to assess the stability and accuracy of the employed models. From the obtained results we found that both median and Dirichlet are the most accurate models over dense and sparse datasets respectively. Also we found that the BetaDR are most stable model across all evaluation measures. Finally, the Fuzzy and LQ were the worst models. From these results we can figure out that while the top three ranked models: median, BetaDR and Dirichlet produce relatively accurate and stable results we recommend using median because it has the simplest implementation among three models, and does not consume cost when running. On the other hand, we recommend to use the median model for products with many ratings and using the Dirichlet model when the products have few number of ratings.

ACKNOWLEDGMENT

The author is grateful to the Applied Science Private University, Amman, Jordan, for the financial support granted to cover the publication fee of this research article.

REFERENCES

[1] Abdel-Hafez, A., Xu, Y. Exploiting the beta distribution-based reputation model in recommender system. Paper presented at the 28th Australasian Joint Conference on Artificial Intelligence, 2015, pp. 1-13.

[2] Bharadwaj, K. K., Al-Shamri, M. Y. H. Fuzzy computational models for trust and reputation systems. Electronic Commerce Research and Applications, 2009, 8(1), 37-47.

[3] F. Maxwell Harper, Joseph A. Konstan. The MovieLens Datasets: History and Context. ACM Transactions on Interactive Intelligent Systems 2015, 5, 4, Article 19.

[4] Jøsang, A., Ismail, R., & Boyd, C. A survey of trust and reputation systems for online service provision. Decision Support Systems, 2007, 43(2), 618-644.

[5] Jøsang, A., Haller, J. Dirichlet reputation systems. In The 2nd International Conference on Availability, Reliability and Security (ARES), 2007, 112-119.

[6] Jøsang, A., Ismail, R. The beta reputation system. In Proceedings of the 15th Bled Electronic Commerce Conference, 2002, 41-55.

[7] Lauw, H. W., Lim, E.-P., Wang, K. Quality and Leniency in Online Collaborative Rating Systems. ACM Transactions on the Web (TWEB), 2012, 6(1).

[8] Leberknight, C. S., Sen, S., Chiang, M. On the Volatility of Online Ratings: An Empirical Study E-Life: Web-Enabled Convergence of Commerce, Work, and Social Life 2012, Vol. 108, pp. 77-86.

[9] Liu, S., Yu, H., Miao, C., Kot, A. C. A fuzzy logic based reputation model against unfair ratings. In Proceedings of the 2013 International Conference on Autonomous Agents and Multi-Agent Systems, 2013, 821-828,

[10] Malik, Z., Bouguettaya, A. Rateweb: Reputation assessment for trust establishment among web services. The International Journal on Very Large Data Bases, 2009, 18(4), 885-911.

[11] Riggs, T., Wilensky, R. An algorithm for automated rating of reviewers. In Proceedings of the 1st ACM/IEEE-CS Joint Conference on Digital Libraries, Roanoke, Virginia, USA, 2001, 381-387, ACM.

[12] Cho, J., Kwiseok K., and Yongtae P.. "Q-rater: A collaborative reputation system based on source credibility theory." Expert Systems with Applications,2009, 36(2), 3751-3760.

[13] Hermoso, R., Roberto, C., Maria ,F. From blurry numbers to clear preferences: A mechanism to extract reputation in social networks. Expert Systems with Applications, 2014, 41(5): 2269-2285.

[14] Chiu, D. K. W., Ho-Fung, L., and Ka-Man, L. On the making of service recommendations: An action theory based on utility, reputation, and risk attitude. Expert Systems with Applications, 2009, 36(2): 3293-3301.

[15] Garcin, F., Faltings, B., Jurca, R. Aggregating reputation feedback. In Proceedings of the 1st International Conference on Reputation: Theory and Technology, 2009, 62-67.

[16] Abdel-Hafez, A., Xu, Y., Jøsang, A. A normal-distribution based rating aggregation method for generating product reputations. In Web Intelligence, 2015, 13(1), pp. 43-51.