Online Reputation Model Using Moving Window

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

***Abstract*— Users are increasingly dependent on decision tools to facilitate their transactions on the internet. Reputation models offer a solution to the users in supporting their purchase decisions. The reputation model takes product ratings as input and produces product quality as score. Most existing reputation models use naïve average method or weighted average method to aggregate ratings. Naïve average method is unstable when there exist a clear trend in the ratings sequence. Also, the weighted methods are influenced by unfair and malicious ratings. This paper introduces a new simple reputation model that aggregates ratings based on the concept of moving window. This approach enables us to study variability of ratings over time which allows us to investigate the trend of ratings and account for sudden changes in ratings trend. The window size can be defined by either number of ratings or duration. The proposed model has been validated against stat-of-art reputation models using Mean Absolute Error and Kendall tau correlation.**

Keywords*— *Reputation Model, Moving Window, Ratings Aggregation Method, E-Commerce.

I. Introduction

E-commerce and mobile commerce systems are increasingly growing in the last two decades which resulted in emergence of new technologies and services [1][6][10]. Therefore, the internet turns into the most common workspace for performing our transactions such as selling and purchasing goods. When users usually want to buy goods on the internet they encounter many similar products, which make the selection process among them is relatively difficult. Accordingly, the users are looking for effective methods that facilitates their decisions. Therefore an accurate and reliable reputation system has moved from novelty and convince to necessity. Almost all B2C and C2C websites ask users to provide ratings and reviews after any successful transaction [1]. For example, eBay users can rate each other, while other review websites, user can rate other reviews as helpful or not. These ratings can serve as decision support tool for other people. It is widely acknowledged that the virtual trust can be achieved when sufficient number of similar ratings are received for a specific product [7]. But in fact, online ratings suffer from many challenges such as unfair ratings, malicious ratings and biased ratings.

Ratings aggregation method is the key part of reputation model which is responsible for calculating the product quality. Many online rating systems use today the simplest method for aggregating rating which is the naïve average method. This method is straightforward, but it cannot discover the trends emerging from recent ratings [10]. To illustrate that, let us have a look at Figure 1 which illustrates that there is a clear trend in ratings where recent ratings are lower than old ratings. So there is no doubt that using naïve average method will fail to inform users with such clear trend. Figure 2 illustrates a different scenario where there is no clear trend in the ratings and there are many sudden changes. In this case, such product should not be given high score. Other studies focused on using rating aging functions to discount old rating [12]. While this approach is theoretically realistic, it might ignore some important hidden knowledge in old ratings, and cannot ensure that the most recent ratings are more informative. Other studies investigated the effect of unfair and malicious ratings on the aggregation process [4][9][11]. They use weighted average method to calculate the product score, where weight can be often the user reputation [2][11]. These approaches have been widely investigated, but most of them focus on one aspect of user reputation, whereas some of them are not working efficient with sparse datasets (i.e. products with few ratings).

None of the previous studies attempted to investigate the variability of ratings over time (i.e. following chronological order of ratings). As explained early, product ratings usually have unclear trend and there are frequently sudden changes in ratings sequence. That is, there is desperate need to involve the variability of ratings in computing product score. This paper uses the concept of moving window to analyze the variability of ratings at specified intervals and then reflect that during aggregation of ratings. The window can be imagine as subset of ratings with a predefined size. The window size is defined by either fixed number of ratings or duration. The big issue is to ensure that there is enough number of ratings for the products with few ratings. The procedure of moving window technique works as follows: Each time the window is shifted one step forward, the variability of window is measured by using statistical variance. This step allows us to see the effect of sudden changes from neighbor's window and detect the ratings that makes sudden changes, these ratings are called unfair ratings.

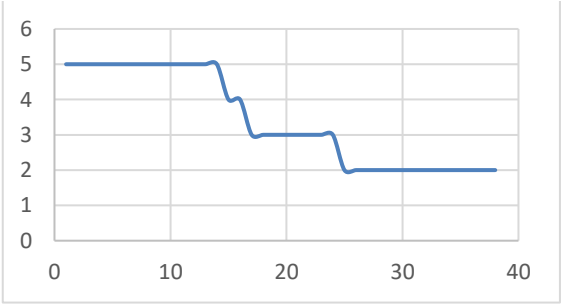

Figure 1. Clear ratings trend over time.

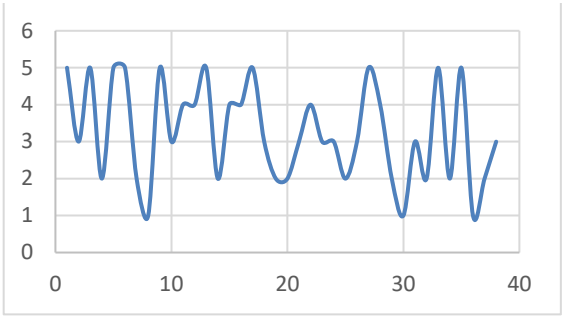

Figure 2. Unclear ratings trend over time.

This paper is structured as follows: section two presents the related works. Section three introduces idea of moving window. Section four presents the proposed model. Section five discusses the datasets. Section six presents evaluation measures. Section seven discusses the results. Section eight contains the conclusions.

II. Related Work

Naïve averaging method is the common method used for aggregating ratings on electronic social and commerce systems [10]. This method is not informative as it cannot discover recent trend in ratings and easily influenced by unfair ratings [10]. In turns, the weighted average method seems more effective as it can take into consideration the quality of users who made ratings. Many studies confirmed that all received ratings for a product are not equal because of many factors such as ratings age, and user credibility, reliability and confidence [4][11]. Therefore many reputation systems attempt to involve one or more factors in computing weight for potential users. Josang et al. [6][7] stated that the ratings age is a good indicator of the importance of ratings. They showed that both linear and nonlinear aging discount function can be used through weighted average method. However, this approach requires determination of decay function parameter such as age unite (i.e. day, week, month and year), which it needs professional experts. On the other hand, another study suggested using the number of past transaction instead of ratings age [12]. Leberknight et al. [10] reveled that recent ratings should have higher weight than old ratings, and the reputation model should take that as discounting factor during ratings aggregation process. They proposed a model that divides rating into number of non-overlapping equal subsets, and then investigate the volatility in each subset with respect to the near subset. Finally the variabilities in all subsets are fused together through discounting function that is used later to compute product score.

Apart of ratings age, other studies investigate the user data in order to discover some important factors such as user reliability, credibility and confidence. In this regard, Riggs et al. [13] suggest including user reliability as weight during rating aggregation process. User reliability means that his provided ratings are very close to the global agreement for all products he rated. Likewise, Lauw et al. [9] studied the leniency ad strictness of users in providing ratings. Lenient user are those who frequently provide positive ratings regardless the actual product quality. Strict users are those who frequently provide negative ratings regardless the actual product quality. Computing leniency of users are quite similar to computing reliability, but the difference is in computing final product score. Jøsang et al. [7] introduced a reputation model based on multinomial Dirichlet probability distribution. Bharadwaj et al. [3] developed some new metrics based on work of Jøsang et al. [7] and using fuzzy logic to compute trust and reputation of product. Cho et al. [4] proposed three factors that assess the user reliability. These factors are user expertise in a specific category, user trustworthiness, and co-orientation. These factors are fused together using either arithmetic mean, harmonic mean, or multiplication. In the same direction, Liu et al. [11] proposed three factors to address the problem of unfair ratings. These factors are fused together using fuzzy logic. The model has been validated using single and multiple attacks procedures.

III. Moving Window

As explained in the introduction section, when the ratings trend is clear the naïve average method cannot discover recent trend in the ratings, analyzing the variability in ratings becomes a must. Moving window is a mechanism similar to moving average method that is used with time series to analyze and find trends in the data. The basic idea of moving window is to analyze the variability of ratings by creating series of variability scores for different windows. The conventional way to measure the variability of ratings inside a window is to use statistical variance method. However, other statistical methods can be also applied. An important question is how to find the appropriate size of window. In practice, there are two approaches to specify the window size, the first approach is to use fixed size window, while another approach is based on duration. The first approach ensures that window size does not change while the window is moving and guarantee the availability of ratings in each window. In contrast, the second approach does not guarantee that all windows have equal number of ratings, but it can provide meaningful interruption for discounting function. Furthermore, this approach does not guarantee that there exist enough number of ratings in each window. In this case the windows without ratings are ignored. In summary, the moving window works as follows: given a series of ratings, the moving window starts with subset of rating, then in each round the window is shifted forward by excluding the first element in the window and added the next rating to the window.

IV. The Prposed Model

In this section the proposed model is described, first, how to measure the variability for each window is described. Then how to reflect the obtained variances on the ratings as discounting factor. Finally the aggregation method is described. Rating variability is calculated based on finding the statistical variance among the ratings in each window. The window size can be determined based on either fixed number of ratings, or duration. In case of duration, one might decide to use week unite as window size, while others might use different time unite based on the timespan of ratings. It is important to note here that window size should not be empty or having small number of

rating as it affects the accuracy of computing ratings variance. In case of fixed number, the chosen size should be reasonable to correctly find variance in ratings. Actually 10 ratings in each window have been chosen. The reason behind this selection is because most online portals displays the most five recent ratings for their visitors.

To illustrate how the moving window works, consider a product has the following ratings sequence ordered chronologically based on the receiving time: $R = \{4,3,5,4,4,5,5,3,4,5\}$ and consider window size is 4. The first window will include {4, 3, 5, 4}, the second window will include {3, 5, 4, 4} and so forth. The last window will include {5, 3, 4, 5}. According to this scenario there will 7 possible windows. The corresponding variance for possible windows are: 0.6667, 0.6667, 0.3333, 0.3333, 0.9167, 0.9167, and 0.9167 respectively. It is clear from the values that there is high variance at the beginning of ratings then that variance began to decrease by time then suddenly rise up to confirm that there is a trouble at latent ratings. Therefore the ratings with large variances should be discounted as it affects the product reputation score. In this example, the ratings in the second and third window should be given greater weight than others.

To reflect the fluctuation of variances as rating weight, a function that maps the obtained variance for each window to a discounting factor is used. The hypothesis of our discounting function is that the window with small variance is given greater weight than other windows of higher variance. In other words, if the variance of the window is very small then greater weight is given to that window. In contrast, if the variance is high then little weight is given to that window. Finally, since each rating may belong to more than window the final discount rate is the average of all discounting rates obtained. The discounting function used in this paper that satisfy our hypothesis is sigmoid function as shown in Equation 1 and Figure 3.

$$f = \frac{1}{1 + e^{-\alpha(var-\lambda)}} \tag{1}$$

Where λ is the variance of the whole product ratings. var is the variance of a particular window. α is a scaling constant to make f is close to 1 when $var \simeq 0$, and can be used to control the sensitivity of the discounting function to the rating fluctuation.

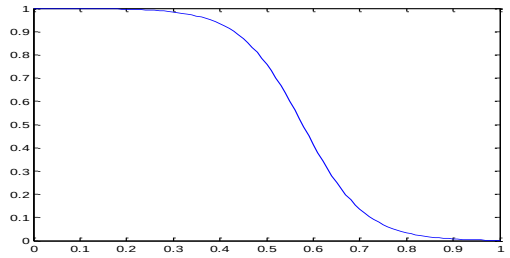

Figure 3. Sigmoid discounting function with $\lambda = 0.57$ and $\alpha = 15$

The aggregation function use weighted average method, where weights are computed from the discounting factor. For each possible window the discount factor is computed. Some ratings receive one discounting value, while other may receive more than one discounting value based on the windows that they belong. In this case the average of these discounting ratios is taken. Finally the obtained ratios are normalized to work as weight for the ratings during aggregation process. Equation 2 shows the weight normalization process. In case of the number of ratings was quite small (i.e. less than or equal to windows size) then only the naïve average of those ratings is considered.

$$w_i = \frac{f_i}{\sum_i^n f_i} \tag{2}$$

$$q = \sum_{i=}^{n} w_i \times r_i \tag{3}$$

Where q is the generated score for a product.

V. DATASETS

Three benchmark datasets that come from one source are used to validate the proposed reputation model. These datasets come from MovieLens data repository [5]. The first dataset called 100K which contains 943 users with 1682 movies and total of 100,000 ratings. The second dataset called 1M which contains 6040 users and 3706 Movies and total of 1,000,209 ratings. The last dataset called 10M which contains 72,000 users and 10,000 movies with total of 1,000,000 ratings. These datasets have been chosen because they are widely used for validating reputation models and publically available so it facilitates replication studies.

Table 1. Datasets characteristics

Dataset	Users Count	Movies Count	Ratings Count
100K	943	1682	100,000
1M	6040	3706	1,000,209
10M	72,000	10,000	10,000,000

VI. EVALUATION MEASUERS

Evaluation measures are good indicators for the accuracy of the reputation models. Unfortunately, there is no accuracy measures that have been agreed among researchers for evaluating the reputation models, however, the most common measures which are Mean Absolute Error (MAE) and Kendall Similarity [1][3] are used. MAE is an indication to the similarity of the product reputation score to the actual ratings of product as shown in Equation 4. Accurate reputation score is the one with very small MAE, close to zero. Kendall similarity measures the correlation between rankings of two lists. It enables us to check if there is agreement between two rankings. The similarity degree falls between -1 (typical disagreement) and +1 (typical agreement). In our case, the good results are achieved when two list have different rankings which confirms that both reputation models are different.

$$MAE = \frac{1}{m}\sum_{j=1}^{m}\frac{\sum_{i=1}^{n}(r_{ji} - q_j)}{n} \quad (4)$$

Where q_j is the generated score for item j. m is the number of items in the testing data. n is the number of ratings for j^{th} item in the testing data.

VII. Results

In practice, the accuracy of any reputation model is usually measured by how helpful were the generated scores to the users in taking right decisions. Unfortunately, this measure is difficult to be obtained because most users do not leave their feedbacks and degree of satisfaction from the given ratings. To avoid this pitfall, it has been suggested to compare between our model and some common reputation models using MAE and Kendall similarity. To successfully perform the comparison among models 10-Fold cross validation is used where the dataset is divided into 10 different training and testing datasets. The training dataset is used to train the model and the testing dataset is used to validate the model. It is important to mention that the original dataset is divided based on the users Ids. In each iteration the MAE is calculated, and the overall MAE is averaged at the end. The reputation models that have been used for comparison purposes are: Naïve average, LQ [9], Dirichlet [7], and Fuzzy model [3]. As discussed in literature, there are two approaches to find the appropriate size of the window: fixed size and duration. In terms of duration it has been preferred to use one month as window size because the online products are not often changed monthly. For fixed window size the 10 ratings have been chosen because most online stores display recent 10 ratings on their websites. The results of comparisons as shown in Table 3. It can be noticed that the moving average in general produces better results over all small and large datasets. Although the difference is not so significant, but it can confirm that our moving window model has the capability to deal with dense and few ratings. Particularly, when comparing between Moving window with fixed size and duration, it can be noticed that using fixed size is slightly better because this approach ensures at least there are sufficient ratings for each window. In practice, some products are not popular so they are not rated regularly, thus it is very hard to ensure that there exist sufficient number of rating in each window.

Table 2. Mean Absolute Error Results

Dataset	Moving Window (fixed size)	Moving Window (duration)	Average	LQ	Dirichlet	Fuzzy
100K	0.7971	0.8106	0.905	1.02	0.898	0.916
1M	0.7891	0.7969	0.841	0.96	0.841	0.848
10M	0.7269	0.7350	0.791	0.92	0.776	0.795

Furthermore, the Kendall similarity between two ranked lists are computed for all comparisons between Moving window (with fixed window size) and each reputation model. To investigate the sensitivity of this analysis, the similarity is computed over a specified percentage of the top ranked products. 10%, 20%....100% have been chosen as threshold points. The main objective of this analysis is that the users are usually concerned about top products, and to confirm that our model produces relatively different list of ranked products from other models. Figures 4, 5 and 6 show the trend for Kendall similarity between our moving window model and each reputation model over each dataset individually. The general trends confirm that our model produce likely different list of ranked product than other models, then this trend decline towards nearly no similarity between them after using 10% top products. This fact suggests that our model and other models generating different lists of top ranked products. Figure 3 shows that that there is slight positive agreement between Moving window and Dirichlet at top 1%, then this degree began to drop when more top products are added to the comparison. Figure 4 shows relatively similar trend to Figure 3, but this time the average method has positive agreement with our moving window model until 10% then began to drop. Figure 5 has quite similar trend to Figure 3. In conclusion, it can be noticed that our model generates different list of top products which confirms our model is different than previous models and show better accuracy as confirmed by MAE.

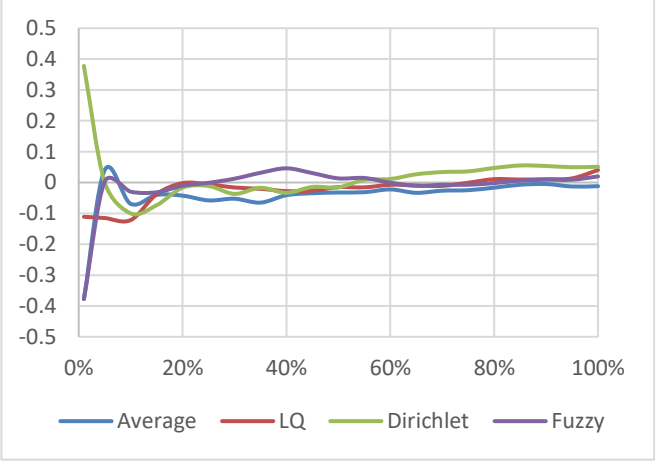

Figure 4. Kendall similarity between Moving window and each compared method over 100K dataset

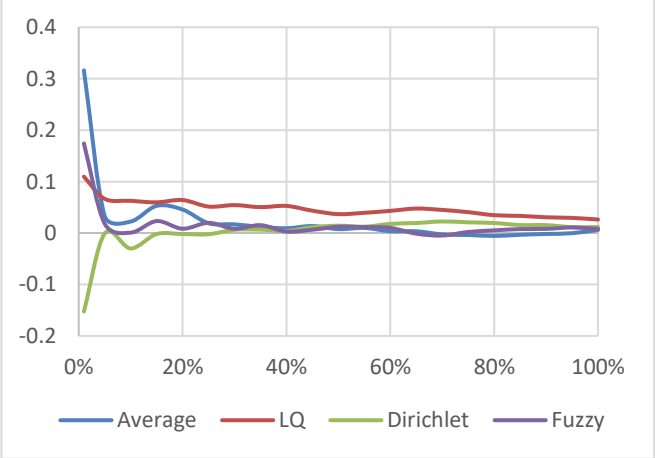

Figure 5. Kendall similarity between Moving window and each

compared method over 1M dataset

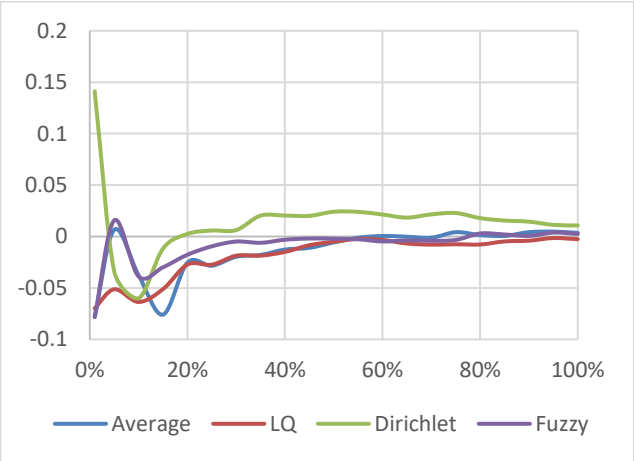

Figure 6. Kendall similarity between Moving window and each compared method over 10M dataset

VIII. Conclusions

Understanding the variability in the ratings over time is an important step towards improving ratings aggregation procedures. When ratings have clear trends, the naïve methods are unstable and cannot discover that trend. But, in reality, product quality are changing over time so the ratings received for that product are likely to change accordingly. Therefore, in this paper a model that can capture the variability in the ratings over time is proposed which can reflect that on the ratings aggregation procedure. The validations over dense and sparse datasets showed that our model produces relatively better accuracy than other previous reputation models. Also, our model generates different list of top ranked products than other models which confirms that our model is significantly different.

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. Australasian Joint Conference on Artificial Intelligence (pp. 1-13). Springer International Publishing, 1-13, 2015.

[2] Aringhieri, R., Damiani, E., Vimercati, D., De Capitani, S., Paraboschi, S., Samarati, P. Fuzzy techniques for trust and reputation management in anonymous peer-to-peer systems. Journal of the American Society for Information Science and Technology, 57(4), 528-537, 2006.

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

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

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

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

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

[8] Kendall, M. G. Rank correlation methods. London, UK: Charles Griffin & Company Limited, 1948.

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

[10] 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 (Vol. 108, pp. 77-86): Springer Berlin Heidelberg, 2012.

[11] 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, 821-828, International Foundation for Autonomous Agents and Multiagent Systems, 2013.

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

[13] 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, 381-387, ACM, 2001.

[14] Tajeddine, A., Kayssi, A., Chehab, A., Artail, H. Fuzzy reputation-based trust model. Applied Soft Computing, 11(1), 345-355, 2011.