Online Reputation Model Using Multiple Quality Factors

Mohammad Azzeh[#], Mohammad Hijjawi[*1], Ahmad M. Altamimi[*2]

[#]*Department of Software Engineering, Applied Science Private University, Alarab street, Amman, Jordan*
E-mail: m.y.azzeh@asu.edu.jo

[*]*Department of Computer Science, Applied Science Private University, Alarab street, Amman, Jordan*
E-mail: [1]hijjawi@asu.edu.jo; [2]a_altamimi@asu.edu.jo

Abstract— **Users on the internet are looking for ways to minimize their experiences in performing online transactions. Reputation systems as a decision support tool are trying to facilitate online transactions. However, many reputation systems use Naïve methods to compute the reputation of an item. These methods are unstable when there is sparsity in the ratings. Also, they cannot discover trends emerging from recent ratings. Other methods, which use weighted average or probabilistic model, usually focus on one aspect of the reviewer ratings. Even though models that combine multiple factors often accomplish that through an arbitrary set of weights. This research study looks at various aspects of reviewers' ratings and proposes a new reputation model that attempts to assess the reviewer reputation by combining four factors through a Fuzzy model. These weights are then involved in computing the item reputation. The proposed reputation model has been validated against state-of-art reputation models and presented significant accuracy regarding Mean Absolute Errors (MAE) and Kendall correlation. The proposed reputation model also works well with the sparse and dense dataset.**

Keywords— **online rating system; reputation model; rating systems; Fuzzy model; aggregation methods.**

I. INTRODUCTION

The massive amount of online transactions is taken place daily on the internet [1]. Many of them are influenced by the opinions and feedback provided by other people. The feedback can be found in two forms: ratings and reviews, which are considered the main sources of item quality information [2], [3]. The quality of the item is usually assessed through a reputation system that is used to process ratings. A crucial part of that system is the rating aggregation method. The outcome of the reputation system is the item quality represented as numeric value or stars in some systems [2], [4].

Using reputation systems is increasingly observed because they are free, widely available, easy to reach, and can facilitate user decision [5]–[7]. However, the accuracy of reputation systems has always been a concern for many users because it reflects global opinion about an item. The primary concern in reputation systems is how to aggregate item ratings in order to reflect the quality of items. The fundamental approach is to use Naïve methods to obtain a reputation score [8]. These methods are rather simple, but they are not efficient because they do not consider the reliability and quality of reviewers or even the popularity of an item. It also cannot discover a trend emerging from recent

reviewer ratings [1]. Other methods attempt to use more sophisticated approaches such as probabilistic models and Fuzzy logic [9], [10]. These approaches presented good accuracy, but they depend heavily on threshold points specified by experts. On the other hand, many reputation models use the weighted average to compute the item reputation score, where the weights may be the reviewers' reputation, reliability [11], leniency of reviewer [12], time of rating [Abdel-Hafez, 2015b], or difference between new rating and current item score [13]. This approach desperately needs evaluating rating quality of reviewer to calculate reviewer weight that will be used during aggregation process based on weighted mean method.

Most of the weighted methods focus on one side of reviewers' reputation, for instance, Lenient-Quality (LQ) model computes the weight based on reviewer leniency and strict in providing ratings [12]. On the other hand, the models that assess multiple aspects usually use the sophisticated discounting function, and subject to thresholds set by the human expert [14], [15]. The second issue is that the weighted average methods do not consider the fluctuation in both reviewer reliability and time between the reviewer's rating and first rating of that item, in addition to the reviewer experience comparing with other reviewers. Furthermore, none of the current weighted average models take the reviewer's tendency in rating items (i.e., positive or

negative ratings of a reviewer). Therefore, this paper presents a new reputation model that can assess the reviewer's reputation by four factors that are combined using a Mamdani Fuzzy model. These factors have been chosen carefully to reflect the reviewer's reputation as weight. These factors are 1) average time fluctuation, 2) reviewer reliability, 3) reviewer behavior or tendency and 4) reviewer experience. More detailed about these factors can be found in section II.*b*.

The primary symbols that are used throughout this paper are:

- $U = \{u_1, u_2, u_3, \dots, u_n\}$ is the set of all reviewers who rated at least one item.
- $E = \{e_1, e_2, e_3, \dots, e_m\}$ is the set of all items that have been rated by at least one reviewer.
- k is the maximum rating level.
- R is a reviewer rating matrix defined as mapping $R = U \times E \to [0, k]$. If reviewer u_i has rated the item e_j then $R(u_i, e_j) = r$ such that $1 < r \ is \le k$, otherwise $R(u_i, e_j) = 0$.
- $|e_j|$ is the number of reviewers who rated item j.
- $|u_i|$ is nua mber of items reviewed by a reviewer i.
- $\overline{R(e_j)}$ is the mean of ratings for itethe m e_j.
- $\bar{R}$ is the mean of all ratings in the dataset.

In literature, many reputation models have been investigated [16], [17], some of these models are being used by popular online applications such as eBay, Amazon, IMDb…etc. However, no agreed method can effectively aggregate the reputation score in a way that reflects the accurate item quality. The most common approach is to use Naïve methods as shown in Equation 1. Garcin et al. [8] examined the common rating aggregation methods such as mean, median and mode, and concluded that the median is the accurate and consistent one. Other authors reported that the Naïve methods are biased, easily influenced by unfair ratings, and cannot discover hidden trends in the recent ratings [1], [14]. IMDb proposed a model that uses true Bayesian estimation as shown in Equation 2, where MinR is the minimum number of rating count required to appear on the top 250.

$$score_j = \frac{1}{|e_j|} \sum_{i=1}^{|e_j|} R(u_i, e_j) \tag{1}$$

$$IMDb = \frac{|e_j|}{(|e_j| + MinR)} \times \overline{R(e_j)} + \frac{MinR}{(|e_j| + MinR)} \times \bar{R} \tag{2}$$

In literature, the reputation models are divided into three main categories: probabilistic models, weighted average models, Fuzzy models. Amongst them, the weighted average method has been extensively studied and examined. The weights in this approach are usually computed based on analyzing past ratings. Josang and Haller [18] used a nonlinear aging function which gives low weights for old ratings and high weights for recent ratings. This approach did not specify exactly which best time unit should be involved in computing the weights. Moreover, the number of past transactions has also been used as weight [7].

Interestingly, none of the above approaches provide stable results when the dataset grows. On the other hand, Leberknight et al. [1] studied the volatility in user rating and reflect that on the user rating. Their model assigns less weight for old rating than current ones. However, this mode does not account for the reviewer's trustworthiness and his credibility in providing accurate ratings.

Other models use reviewer data such as reliability, credibility, and trustworthiness to compute weights [15]. Riggs et al. [13] used the reviewer's reliability as a measure of weight, where high weights are offered to the reviewer with accurate ratin that are very close to the average of items scores. Lauw et al. [12] proposed a model that can compute the leniency of the reviewer based on his rating tendency as shown in Equations 3 and 4. The model used leniency variable l_i to divide the reviewers into two groups (strict and lenient) where $l_i > 0$ indicates that the reviewer is lenient.

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

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

Where q_j is the initial score and usually begins with average of ratings. $\alpha \in [0, 1]$ is a compensation factor specified by expert.

Abdel-Hafez et al. [2] proposed two different reputation models using Normal and Beta distribution function to treat sparse ratings. They assumed that the level with large rating count should receive the high weight. Furthermore, using Beta distribution has the flexibility to change the shape of distribution according to α and β parameters, which are calculated based on methe an and standard distribution of ratings. In order to successfully use their method, it is important to sort the rating ascendingly, then apply Beta distribution function to obtain the weights as shown in Equation 5. Equation 6 is used to evenly divide the space of ratings into $|e_i|$ indexes which can be used in Equation 5 to compute Beta weight. Equation 7 is used to calculate item reputation score.

$$Beta(x_i) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x_i^{\alpha-1} (1 - x_i)^{\beta-1} \tag{5}$$

$$x_i = \frac{0.98 \times i}{|e_i| - 1} + 0.01 \tag{6}$$

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

Where Γ is the gamma function, and w_l is the summation of normalized Beta weight for the target level. Jøsang et al. [19] used Dirichlet probability distribution and multinomial Bayesian probability distribution to treat uncertainty in the items with sparse ratings as shown in Equations 8 and 9.

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

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

where $\overrightarrow{S_y}$ is the score vector of a rating level, $S_y(i)$ is the probability that one agent gives rating i to agent y. $R_y(i)$ is the number of ratings of the level i.

Bharadwaj et al. [11] used the ordered weighted averaging method as shown in Equations 10 to 12. Liu et al. [14] used the Fuzzy logic model to treat unfair testimonies. They used several measurements by combining Fuzzy logic to identify the unfair rating. The output of Fuzzy model is used as discounting weight. The main concern in this method is that the rating's weight is high in most cases especially when the most recent rating has received and the similarity between reviewers and users is also high [2].

$$score_j = \sum_{i=1}^{|e_j|} W_i \times R(u_i, e_j) \quad (10)$$

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

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

To the best of our knowledge, none of the current methods addressed the reviewer reputation by combining our four proposed factors through the Fuzzy model. Also, there is no model can work the overall type of datasets (i.e., sparse or dense) accurately.

The remaining of the paper is structured as follows: Section two describes the Material and methods used in the study including our proposed model and research methodology. Section three presents the results, and finally, Section four ends with a conclusion.

II. MATERIAL AND METHOD

A. Fuzzy Logic Model

Fuzzy Logic Model is a rule-based method that can treat uncertain and imprecise problems. Figure 1 describes the basic representation of Fuzzy Logic Model. The Model consists of four main components. The first component is the Fuzzifier which takes input data as crisp real value and maps them to Fuzzy sets with membership values. Each input variable has a universe of discourse which is divided into some Fuzzy sets represented by Fuzzy membership functions. Each Fuzzy set is a representation of linguistic term, for example, suppose we have a variable called temperature, and this variable is represented by three linguistic terms (low, medium and high), so each scientific term is expressed as a fuzzy set with proper fuzzy membership function. The second part is the Fuzzy rule base which stores all rules that are usually implemented by experts. The rules are usually written using if-then syntax. The key component of the Fuzzy Logic system is the Fuzzy

Inference System that can translate input space into output space using Fuzzy rules. The last part is the defuzzification function that translates the output membership values into quantifiable value.
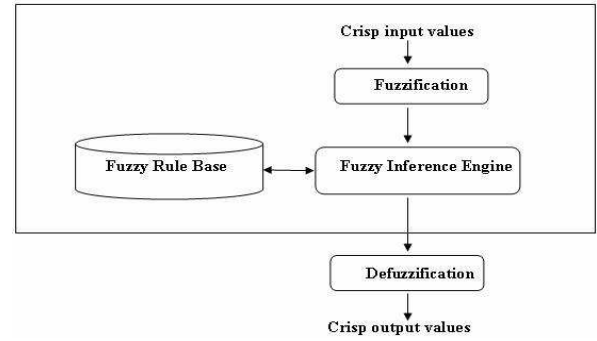

Fig 1. Fuzzy Logic Model

B. The Proposed Model

The proposed model is a weighted average method where weights are calculated through Fuzzy Logic Model. Mainly, we proposed four new factors that measure and reflect the reviewer's reputation. These factors are fused through Mamdani Fuzzy Logic Model whose output is used as a reviewer discounting weight. Using Fuzzy Logic has the advantage of minimizing human decisions and avoiding crisp pre-set thresholds that are used to calculate the discounting weights.

The first factor is the fluctuation in rating timestamp. It is important to study when the reviewer frequently made his/her ratings and if previous ratings of such items might influence he/she. It is very common in online transactions that existing ratings influence reviewers, and they frequently provide rating very close to the average of current ratings. However, in case that he was the first reviewer who rated the item, his/her rating might be more reasonable and reflects the quality of the item. Therefore, for each item assessed by a reviewer, we compute the distance between the most recent timestamp of his rating and the timestamp of the first rating received. If he/she were the first reviewer for that item, then he would receive the high weight.

In contrast, if he/she rates an item with many past ratings, then the weight value is discounted based on the distance between the two timestamps. Finally, we calculate the average for all discounting values across all items as shown in Equation 13. The λ is the fading variable that is used as a discounting factor. In our case we used λ=0.95.

$$f_{i1} = \frac{1}{|u_i|} \sum_{j}^{|u_i|} \lambda^{t_{cj} - t_{1j}} \quad (13)$$

where t_{cj} is the time when item j has been rated by revthe iewer u_i. t_{1j} is the time of the first ratings of item j. Both timestamp parameters are measured in weeks.

The second factor is the fluctuation in reviewer ratings (i.e., reviewer reliability) as shown in Equation 14, which measures the fluctuation over time in providing an accurate rating to the items under assessment. The accurate rating means that the ratings provided by a reviewer should be

2614

close to the average rating of the item under assessment. For each item assessed by a reviewer, the distance between reviewer rating and an average of item ratings is calculated as shown in Equation 15. All distance values are then used with discounting function λ to compute rthe eliability of rthe eviewer as shown in Equation 14. In this equation, if all distance values are very close to zero ,t hen the reviewer would receive a value very close to one.

$$f_{i2} = \frac{1}{|u_i|}\sum_{j}^{|u_i|}\lambda^{\Delta_{ij}} \tag{14}$$

$$\Delta_{ij} = \left| R(u_i, e_j) - \overline{R(e_j)} \right| \tag{15}$$

The third proposed factor is to assess the reviewer's behavior (or tendency) in providing positive and negative ratings. Some reviewers are strict in providing their ratings, so they usually provide negative ratings irrespective of the item quality. In contrast, other reviewers tend to be lenient and usually provide positive ratings. To reflect that tendency on the reviewer's reputation, we first classify the ratings of the reviewer into three classes: positive, neutral and negative. Positive ratings are those with the values 4 and 5 in the rating level, neutral is a set of ratings that have a value of 3, and Negative ratings are those with the values 1 and 2. To compute this factor, in the first step we classify reviewer's items into three groups based on dividing mentioned above criteria. In the second step, we find the ratio between the sum of residuals within the target class and the sum of residuals within the reviewer's rating as shown in Equation 16. The target class is determined based on the rating of the item under assessment; for example, if the rating given by a reviewer for the item under assessment is 5, then we compute Equation 16 for the positive class only.

$$f_{i3} = 1 - \frac{\sum_{k=1}^{L}\left| R(u_i, e_k) - \overline{R(e_k)} \right|}{\sum_{j=1}^{|u_i|}\left| R(u_i, e_j) - \overline{R(e_j)} \right|} \tag{16}$$

Where L is the number of ratings in the target class.

The fourth factor is to assess the reviewer's experience in providing an accurate rating. The reviewer's experience is significant in determining the reviewer's confidence and his ability to provide true ratings. This factor can be assessed by finding the ratio between the number of ratings provided by the reviewer u_i and maximum reviewer ratings in the dataset, as shown in Equation 17.

$$f_{i4} = \frac{|u_i|}{max\{|u_1|, |u_2|, |u_3|, \ldots, |u_n|\}} \tag{17}$$

The four factors are fused using Mamdani Fuzzy Logic model. The outcome is the reviewer reputation as weight value between 0 and 1. Each input factor was denoted by three Fuzzy sets (Low, Medium and High). Whereas, five fuzzy sets denote the output variable (Very Low, Low, Medium, High and Very High) to give more flexibility in computing the weight. Each Fuzzy set is represented by a trapezoidal membership function as shown in Figure 2. Equation 18 defines Fuzzy trapezoidal membership function.

$$\mu(x; a, b, c, d) = max\left(min\left(\frac{x-a}{b-a}, 1, \frac{d-x}{d-c}\right), 0\right) \tag{18}$$

Where a, b, c, and d are Fuzzy trapezoidal membership function parameters. X is a numeric value.
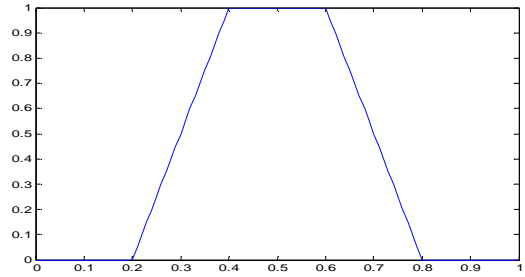

Fig. 2 Symmetric Trapezoidal Membership Function

The universe of discourse of each variable was partitioned equally among all fuzzy sets. Eighty-One Fuzzy rules were carefully constructed to map input space to output space. The full list of the constructed Fuzzy rules is presented in Appendix A.

C. Rating Prediction Procedure

In the previous section, we demonstrate how the prosed factors are computed and fused to produce reviewer weight. However, in this section, we illustrate the rating prediction procedure. Given an item (e_i) from the items repository E we extract set of reviewers who rated this item. Then, for each reviewer (u_i) we compute the four factors f_{i1}, f_{i2}, f_{i3} and f_{i4}. These factors are entered into the Fuzzy Logic Model that is supposed to be constructed as explained in sections 4 and 5. The input of Fuzzy model should be in the following format: $\langle f_{i1}\ f_{i2}\ f_{i3}\ f_{i4}\rangle$. The outcome of the Fuzzy Model (i.e. reviewer weight $w_{,i}$) is calculated through the defuzzification process. The final item score is computed as shown in Equation 19.

$$score_j = \sum_{i=1}^{|e_j|}\frac{w_i \times R(u_i, e_j)}{\sum_{i=1}^{|e_j|} w_i} \tag{19}$$

D. Datasets

To validate the proposed reputation model, we used common stable benchmark datasets called MovieLens [20]. This repository contains several benchmark datasets with a different number of anonymous ratings. For the objective of this research, we used two stable versions of the MovieLens datasets (100K and 1M) to validate the accuracy of the proposed model in the absence and presence of sparse or dense data [20]. Also, we extracted three new datasets from the 1M dataset for the goal of examining the efficiency of our model over the different level of sparsity. The extracted datasets contain four, six and eight reviewer ratings selected randomly from reviewers' ratings to that item. The new datasets are called 1M4, 1M6 and 1M8 respectively. The complete description of these datasets is presented in Table 1.

TABLE I
DATASETS CHARACTERISTICS

Dataset	#Reviewers	#Items	#ratings
100K	943	1682	100.000
1 M	6040	3706	1.000.209
1 M4	6040	920	24.160
1 M6	6040	1286	36.240
1 M8	6040	1625	48.320

E. Evaluation Measures

It is widely acknowledged that there are no agreed evaluation measures among researchers for appraising reputation models. To avoid this pitfall, we used some proposed measures that were commonly used in many previous studies [2], [11]. The first measure is the Mean Absolute Errors (MAE) which evaluates, the closeness of the predicted score to the actual ratings for an item. To compute MAE, we measure, for all test cases, the difference between predicted item ratings and actual ratings as shown in Equation 20. This measure is considered unbiased and can give us true picture about the performance of any reputation models. The model that resulted with minimum MAE is considered the superior one.

$$MAE = \frac{1}{m}\sum_{j=1}^{m}\frac{\sum_{i=1}^{h}\big(R(u_i,e_j) - score_j\big)}{h} \qquad (20)$$

Where $score_j$ is the predicted rating. m is the number of items in the testing data. h is tHe number of ratings for j^{th} itthe em in the testing data.

The second evaluation measure is Kendall Tau coefficient which measures association and its significance between two ranked lists. Kendall Tau coefficient takes values between -1 (perfect strong agreement) and +1 (perfect strong disagreement). The purpose of using this measure is to ensure that the ranking order of top items in our models and another model are entirely different. If both ranking lists are different, then two models produced different items scores and confirmed that our model is significant. Otherwise, two models can generate similar items scores. Therefore there is no novelty in our work.

III. RESULTS AND DISCUSSION

This section presents the results of validation and comparison with state-of-art reputation models. Specifically, we compare our model to seven reputation models: Average, BetaDR [2], Bayesian [19], Dirichlet [18], IMDb, Fuzzy rating [11] and LQ [12]. 10-Fold cross-validation has been used as a validation procedure where the original dataset is divided into ten subsets of training and testing observations. In each validation run, the training reputation model is constructed from 90% of the reviewers' data and validated over testing data that forms 10% of original data. In each iteration, we compute Mean Absolute Error for test ratings. This procedure is repeated ten times to ensure that all observations have been used as testing. The first part of validation includes computing of MAE for each reputation model overall employed datasets. Tables 2 shows the obtained MAE results over sparse and dense datasets.

We can observe that our model presents better accuracy in comparison with other models. Correctly, we can conclude that the proposed model works well over sparse and dense dataset as confirmed by the results. However, our model has been outperformed by the Dirichlet model over one sparse dataset 1M6, but the difference is not that significant. This is because the Dirichlet method can treat uncertainty problem that is inherited from sparse datasets. In general, our model and Dirichlet model are relatively close over sparse datasets, and this confirms the ability of our mode to handle the uncertainty in the sparse datasets. The significant improvement of our model is observed especially over dense datasets (100K and 1M).

Nevertheless, we can notice that both IMDb and our model produce accurate results over large dense datasets. This confirms the superiority of our model on dense datasets. Our model keeps the advantage of both dealing with few ratings and dense ratings. Finally, the LQ results Are relatively the worst results over sparse and dense datasets without any enhancements over Naïve method (arithmetic mean).

The second part of validation is to use Kendall tau coefficient to measure the degree of similarity or association between the items ranking produced by our model and the ranking of the items produced by each one of the comparative models. The objective of this evaluation is to show that our model ranks top items substantially different from other reputation models.

TABLE II
EVALUATION RESULTS USING MEAN ABSOLUTE ERROR RESULTS

Dataset	Our Model	Average	BetaDR	Bayesian	Dirichlet	IMDb	Fuzzy Rating	LQ
100K	**0.8154**	0.8737	0.8607	0.8816	0.8677	0.8805	0.8848	0.9906
1M	**0.7763**	0.8348	0.8275	0.8365	0.8335	0.7770	0.8427	0.9527
1M4	**0.7145**	0.7570	0.7556	0.7374	0.7585	0.7944	0.7625	0.8407
1M6	0.7861	0.8020	0.7928	0.8017	**0.7747**	0.8022	0.8090	0.8975
1M8	**0.7791**	0.8123	0.8016	0.8020	0.7848	0.8098	0.8163	0.9090

For each comparison, we run the Kendall tau coefficient 21 times on the ranking, wherein each time we take a specific percentage of top items in both models. In the first round, we compute Kendall tau coefficient over 1% of the top items in both models. This step is significant to see the similarity between top items that are usually used by recommendation systems for other users as suggestions. The good results should include minimum similarity value. For other 20 rounds, we take top 5%, 10%, 15%… 100% respectively. This procedure is repeated for all dense and sparse datasets which enable us to understand the trend of similarity between the ranking of top items in our model and other reputation models. The outcome of each comparison over all datasets is shown visually in Figures 3 to 7. The horizontal axis represents the percentage of top items. The vertical axis represents the value of Kendall tau coefficient between our model and models list in each figure. The general conclusion that can be revealed from these figures is that there is a common trend in all comparisons such that they started either in quite perfect agreement or disagreement and began to decline until they reach a steady

state which is very close to zero. For dense datasets (i.e., K100 and M1) we can notice that our model produces quite similar top item list to BetaDR when we look at the top 1% and 5% of the items. For K100 dataset as shown in Figure 3, it is noticed that our model ranks 1% and 5% of top item quite similarly to Bayesian and Fuzzy rating models. However, the similarity degree began to decline after using top 10%.

In contrast, our model ranks the list of top items differently from other reputation models. Likewise, we can see the same trend over M1 as shown in Figure 4, but with one exception for BetaDR which ranks top items differently from our models at various percentages of top items. Notably, our model and both the Dirichlet model and average rank top items differently, which indicates that our model is more accurate as confirmed by MAE. For sparse datasets, we notice that our model and both Bayesian and Fuzzy ratings rank only top 1% and 5% items similarly, but they decline after using 10%, which confirms that ranking lists are independent of each other. For other comparisons over sparse datasets, we can notice that our model ranks top items differently from other reputation models as shown in Figures 5, 6 and 7. In summary, we can figure out that the ranking order of the top 10% of item list generated by our model is relatively different from other reputation models, either over sparse or dense datasets.
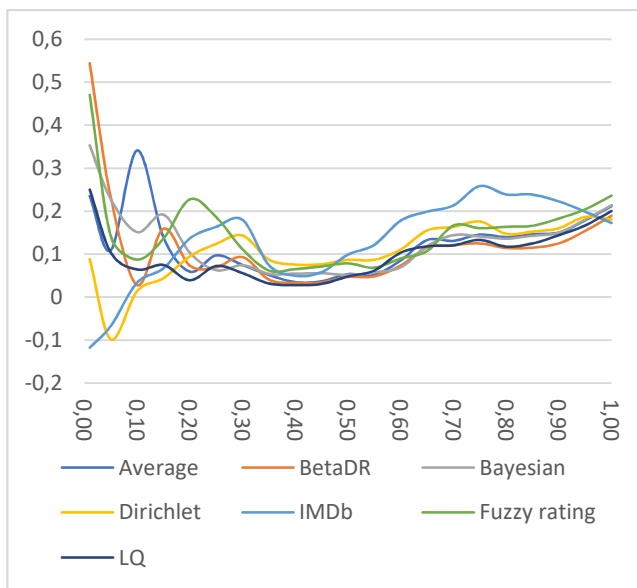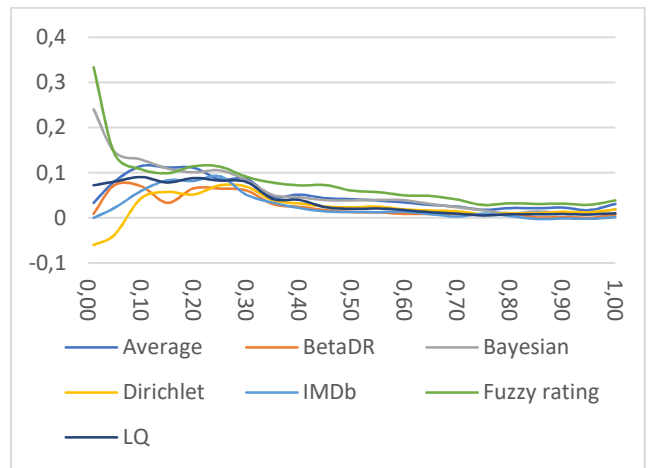

Fig. 4 Kendall tau coefficient comparisons over M1 dataset
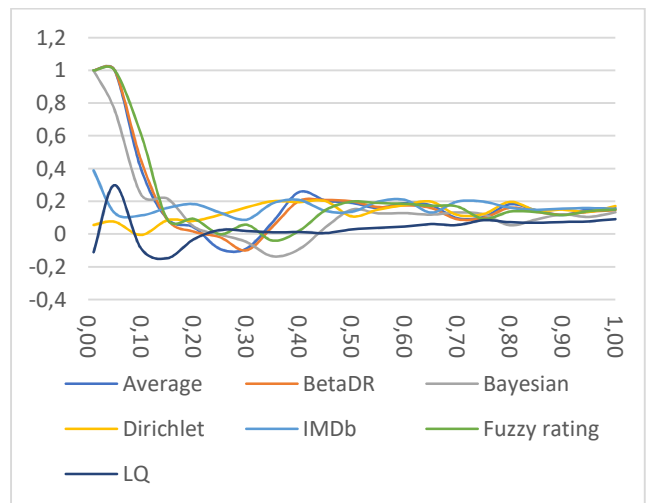

Fig. 5 Kendall tau coefficient comparisons over the 1M4 dataset

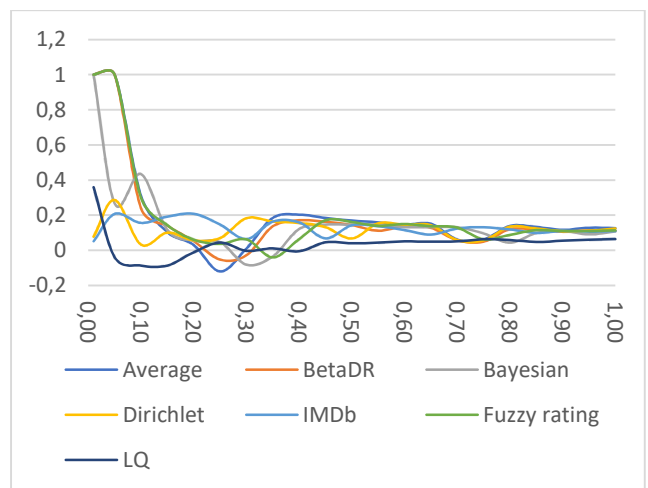Fig. 3 Kendall tau coefficient comparisons over K100 dataset

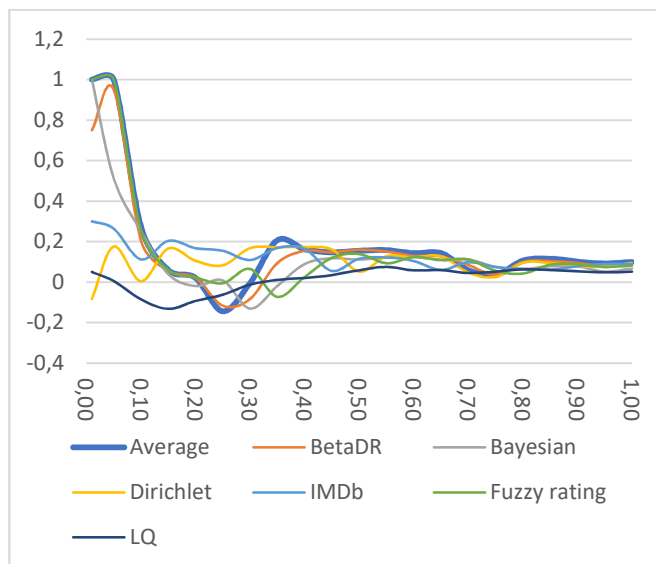Fig. 6 Kendall tau coefficient comparisons over 1M6 dataset

Fig. 7 Kendall tau coefficient comparisons over the 1M8 dataset

IV. CONCLUSION

It is widely acknowledged that malicious and unfair ratings easily influence rating aggregation process so the reviewer reputation should be involved in this task. This paper introduces a new reputation model that uses four new factors fused through a Fuzzy Logic Model to compute the reviewer's reputation. This value is then used as a weight in computing the item quality. Each factor measures one aspect of reviewer reputation, which enables us to assess his/her ratings and their influence on item quality. The proposed model solved some limitations in previous models and showed good accuracy regarding MAE. The proposed model provides great enhancement for the accuracy of rating aggregation process over sparse and dense datasets. The comparisons between our model and other legacy reputation modes using Kendall tau coefficient showed that, in most cases, our model ranks the top items differently from other reputation models which indicates the significant improvements of our model. There are some exceptions especially over sparse dataset because our model and both Fuzzy rating model and Bayesian model are relatively quite similar at the top 1% and 5% only. In summary, our model and other models are significantly different at greater than or equal to the top 10% of the item list.

ACKNOWLEDGMENT

REFERENCES

[1] 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.

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

[3] Ayday, E., Lee, H., Fekri, F. An iterative algorithm for trust and reputation management. In IEEE International Symposium on Information Theory (ISIT), 2009, 2051-2055.

[4] 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.

[5] 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.

[6] Bedi, P., Sharma, R. Trust based recommender system using ant colony for trust computation. Expert Systems with Applications, 2012, 39(1), 1183-1190.

[7] 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.

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

[9] 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, 2006, 57(4), 528-537.

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

[11] 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.

[12] 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).

[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, 2001, 381-387, ACM.

[14] 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,

[15] 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.

[16] 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.

[17] 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.

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

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

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

APPENDIX A . FUZZY RULES OF OUR PROPOSED MODEL

#	f1	f2	f3	f4	Output	#	f1	f2	f3	f4	Output	#	f1	f2	f3	f4	Output
1	L	L	L	L	VL	28	M	L	L	L	VL	55	H	L	L	L	L
2	L	L	L	M	VL	29	M	L	L	M	L	56	H	L	L	M	M
3	L	L	L	H	L	30	M	L	L	H	M	57	H	L	L	H	M
4	L	L	M	L	VL	31	M	L	M	L	L	58	H	L	M	L	M
5	L	L	M	M	L	32	M	L	M	M	M	59	H	L	M	M	M
6	L	L	M	H	M	33	M	L	M	H	M	60	H	L	M	H	M
7	L	L	H	L	L	34	M	L	H	L	M	61	H	L	H	L	M
8	L	L	H	M	M	35	M	L	H	M	M	62	H	L	H	M	M
9	L	L	H	H	M	36	M	L	H	H	M	63	H	L	H	H	H
10	L	M	L	L	VL	37	M	M	L	L	L	64	H	M	L	L	M
11	L	M	L	M	L	38	M	M	L	M	M	65	H	M	L	M	M
12	L	M	L	H	L	39	M	M	L	H	M	66	H	M	L	H	M
13	L	M	M	L	L	40	M	M	M	L	M	67	H	M	M	L	M
14	L	M	M	M	M	41	M	M	M	M	M	68	H	M	M	M	H
15	L	M	M	H	M	42	M	M	M	H	M	69	H	M	M	H	VH
16	L	M	H	L	M	43	M	M	H	L	M	70	H	M	H	L	H
17	L	M	H	M	M	44	M	M	H	M	M	71	H	M	H	M	H
18	L	M	H	H	M	45	M	M	H	H	H	72	H	M	H	H	VH
19	L	H	L	L	L	46	M	H	L	L	M	73	H	H	L	L	M
20	L	H	L	M	M	47	M	H	L	M	M	74	H	H	L	M	H
21	L	H	L	H	M	48	M	H	L	H	M	75	H	H	L	H	H
22	L	H	M	L	M	49	M	H	M	L	M	76	H	H	M	L	H
23	L	H	M	M	M	50	M	H	M	M	M	77	H	H	M	M	H
24	L	H	M	H	M	51	M	H	M	H	H	78	H	H	M	H	VH
25	L	H	H	L	M	52	M	H	H	L	H	79	H	H	H	L	H
26	L	H	H	M	H	53	M	H	H	M	H	80	H	H	H	M	VH
27	L	H	H	H	H	54	M	H	H	H	VH	81	H	H	H	H	VH

Abbreviations: **VL**: Very Low   **L**: Low   **M**: Medium   **H**: High   **VH**: Very High