ICIC Express Letters
Volume **11**, Number **2**, February **2017**

ICIC International ©2017 ISSN 1881-803X
pp. 381–389

A RESTAURANT RECOMMENDATION ALGORITHM BASED ON IMPROVED COLLABORATIVE FILTERING

Yancui Shi, Xiankun Zhang, Yiying Zhang and Jianhua Cao

College of Computer Science and Information Engineering
Tianjin University of Science and Technology
No. 1038, Dagu Nanlu, Hexi District, Tianjin 300222, P. R. China
syc@tust.edu.cn

Received July 2016; accepted October 2016

Abstract. *In order to alleviate the sparsity problem of the collaborative filtering (CF) and improve the accuracy of the recommender system, a restaurant recommender system (RRS) based on the improved CF is proposed by analyzing the tags information in dianping.com. Firstly, the mean centering is employed to preprocess the rating in order to reduce the error caused by different personal habits. Secondly, according to the feature of the restaurant and the user rating, an improved null filling method is proposed to alleviate the sparsity of the user-restaurant matrix. And then, when calculating the user weight, the similarity of the user preference and the user trust are considered, which can improve the accuracy of the recommender system. Finally, the experiment is executed in the real data set and the experimental results show that the proposed method can obtain higher accuracy than the existing methods.*

Keywords: Recommender system, Restaurant recommendation, Collaborative filtering, Similarity, Trust

1. **Introduction.** In recent years, the recommender system as a tool of information filtering attracts a large number of scholars to research [1]. Currently, the recommender system has been applied in a variety of multimedia and e-commerce websites, such as Amazon, Google news and Taobao. The recommender system mainly is divided into three categories: the system based on content, based on CF and based on hybrid model [2]. The recommender system based on CF is the most classical and most widely used [3].

With the rapid development of catering industry, how to provide the personalized restaurant for user timely and accurately has become the focus of the current research. Ge et al. considered the tags information of user and restaurant and latent factors, and realized the personalized food recommendation using the improved matrix factorization (MF) algorithm [4,5]. While Kuo et al. recommended the personalized restaurant for user according to the history logs of user booking [6]. Katarya and Verma introduced the psychological factors (e.g., lifestyle, preference), the demographic information (e.g., age, gender) and location information into the food recommendation [7]. Similarly, Zhang et al. analyzed the history information of user dining, demographic characteristics and the restaurant's feature, and recommended the next dining according to the characteristic of each user [8]. However, the above research did not consider the influence of the social relationship and time factor; hence, inspired by the existing research, the proposed method considers the tags of the user and restaurant, and obtains the user trust according to the tags information.

The organization of the paper is as follows: Section 1 introduces the background and the related work; the proposed method is depicted in details in Section 2, including the normalization of the rating, the improved null filling method, the calculation of the user weight and the prediction of the user preference; Section 3 gives the experiments and the analysis of results. The conclusion is shown in the final section.

2. **Proposed Method.**

In the paper, the two-dimensional rating matrix of the user-restaurant ($U \times D$) is employed to represent the user preference. Here, $u_i \in U$ represents the user i; U represents the set of users; $d_j \in D$ represents the restaurant j; D represents the set of the restaurants; $r_{ij} \in (U \times D)$ is the element of the matrix, and represents the rating that u_i rated d_j, and its value is the integer in $0 \sim 5$; $f_k \in F_e$ represents d_j's feature k; F_e represents the set of the selected restaurant features; $C_{i,k}$ represents the set of the feature values that the feature f_k of the restaurant that user i has rated includes; $c_{i,k,l} \in C_{i,k}$ represents a specific feature value.

The proposed method improves the accuracy of RRSS by alleviating the sparsity problem. Firstly, the improved null filling method is employed to calculate the null ratings. And then, when finding the nearest neighbors of the target user, not only does it consider the similarity of user preference, but also it considers the user trust.

2.1. **The normalization of the rating.**

When different users rate the restaurant, even though they have the same recognition, they may also give different ratings. For example, some users are not willing to give the high rating to the favorite restaurant or give the low rating to the disliked restaurant, so the distinction of the ratings is not large. While other users are opposite, and the distinction of the given ratings is very large.

In the recommender system, it needs to predict the preference of the target user according to the ratings of the other users. If the rating is introduced into the recommender system directly, it leads to the low accuracy of recommender system. Because the two users who have different preferences may have the same rating habits, while the rating that the users with similar preferences give may have large gap, when seeking the nearest neighbors of the target user, it is necessary to normalize the rating.

In the paper, the mean centering method is employed to normalize the rating. The formula is as follows:

$$\delta_{ij} = \frac{r_{ij} - \overline{r_{u_i}}}{r_{i,\max} - r_{i,\min}} \tag{1}$$

where $\overline{r_{u_i}}$ represents the average of u_i's rating; $r_{i,\max}$ and $r_{i,\min}$ represent the maximum and minimum respectively; δ_{ij} represents the normalized rating, and it may be positive or negative.

2.2. **The improved null filling method.**

In real life, the number of the restaurant is massive, while the explicit rating given by users in the website is limited. In addition, when there is no commonly rated restaurants, it cannot calculate the similarity between users. The common method to solve this problem is the null filling method [9]. Inspired by the existing research, in the paper, the ratings and the restaurants' features are introduced into RRS and the null filling method is employed to obtain the appropriate fault value. When setting the fault value, the weight of the restaurant's feature is considered.

The weight of the restaurant's feature is determined according to the volatility of the user's rating, and the greater the volatility is, the greater the weight is. The volatility of the feature f_k is as follows:

$$vol_{u_i,f_k} = \frac{\sum\limits_{c_{i,k,l} \in C_{i,k}} \left| n_{c_{i,k,l}} - \overline{n_{f_k}} \right|}{\overline{n_{f_k}}} \tag{2}$$

In the paper, the selected features of the restaurant include the type of restaurant, average consumption and geographical location. $n_{c_{i,k,l}}$ represents the number of the restaurants that u_i has rated and the feature value of those restaurants is $c_{i,k,l}$. $\overline{n_{f_k}}$ is the average of the number that u_i has rated the restaurants with the feature f_k. When the feature is the type of the restaurant, the feature value is the type of the restaurant given in the website. When the feature is average consumption or geographical location, the whole range of value is divided into several small ranges, and each small range represents a category, that is a feature value. The formula that is used to calculate the weight of f_k is as follows:

$$w_{u_i,f_k} = \frac{vol_{u_i,f_k}}{\sum\limits_{f_m \in F_e} vol_{u_i,f_m}} \tag{3}$$

In the improved null filling method, it needs to predict the rating of the restaurant that the user did not rate. Firstly, it needs to select restaurants that have the same feature value with the target restaurant. And then, it needs to calculate the rating according to the rating of those restaurants, the weight of the corresponding feature and the specific feature value of those restaurants. The formula is as follows:

$$\delta_{ij} = \sum_{k=1}^{3} w_{u_i f_k} * \frac{n_{c_{i,k,l,j}}}{\sum\limits_{c_{i,k,r} \in C_{i,k}} n_{c_{i,k,r}}} * \frac{\sum\limits_{d_q \in D_{c_{i,k,l,j}}} \delta_{d_q}}{\left|D_{c_{i,k,l,j}}\right|} \tag{4}$$

where $n_{c_{i,k,l,j}}$ represents the number that u_i has rated the restaurant which has the same feature value with the feature f_k of d_j; $D_{c_{i,k,l,j}}$ represents the set of the restaurants that have the same feature value with the feature f_k of d_j; the restaurant's rating is given in the website and it is the average of all users' ratings, δ_{d_q} represents the normalized rating of d_q.

2.3. **The calculation of the user weight.** When calculating the user weight, the impact of the similarity and trust of user is considered. Firstly, it needs to predict the null value of the users besides the target user using the improved null filling method.

1) the calculation of the user similarity

The user's preference may change over time, and when the user recommends restaurant for the other users, he usually selects the recent visited restaurant. In the recommender system, the user's rating represents the degree that the user is interested in at some time. However, the user's interest will decay, that is the utility of user's rating will decay. Aiming at the forgetting phenomenon, psychologist Ebbinghaus proposed the forgetting function $J(t)$ [10]. The function is as follows:

$$J(t) = \frac{ae^b}{(t+t_0)^c}, \quad a > 0, \quad c > 0, \quad b > 0, \quad t_0 > 0 \tag{5}$$

where t represents the time difference from the rated time to current time, and the unit is day. The values of those variables were given in [10]: $a = 20$, $b = 0.42$, $c = 0.0225$, $t_0 = 0.00255$.

In the paper, the improved Pearson correlation coefficient is employed to measure the similarity between users, and the formula is as follows:

$$\begin{aligned} Sim(u_i,u_j) = & \frac{\sum\limits_{d_k \in D_{ijo}} \left(\delta_{ik}J(t_{ik}) - \overline{\delta_{u_i}}\right)\left(\delta_{jk}J(t_{jk}) - \overline{\delta_{u_j}}\right)}{\sqrt{\sum\limits_{d_k \in D_{ijo}} \left(\delta_{ik}J(t_{ik}) - \overline{\delta_{u_i}}\right)^2 \sum\limits_{d_k \in D_{ijo}} \left(\delta_{jk}J(t_{jk}) - \overline{\delta_{u_j}}\right)^2}} \\ & + \theta * \frac{\sum\limits_{d_k \in D_{ijn}} \left(\delta_{ik}J(t_{ik}) - \overline{\delta_{u_i}}\right)\left(\delta_{jk}J(t_{jk}) - \overline{\delta_{u_j}}\right)}{\sqrt{\sum\limits_{d_k \in D_{ijn}} \left(\delta_{ik}J(t_{ik}) - \overline{\delta_{u_i}}\right)^2 \sum\limits_{d_k \in D_{ijn}} \left(\delta_{jk}J\left(t_{jk}\right) - \overline{\delta_{u_j}}\right)^2}} \end{aligned} \tag{6}$$

where D_{ijo} represents the set of the common restaurants rated by u_i and u_j; D_{ijn} represents the set of the restaurants that has been filled rating for u_j and rated by u_i. Compared with the user's rating, the credibility of the filled rating is lower, so the parameter θ is used to reduce the impact of the filled rating. $J(t_{ik})$ and $J(t_{jk})$ represent the forgetting function of r_{ik} and r_{jk} that u_i and u_j rated the d_k respectively; $\overline{\delta_{u_i}}$ and $\overline{\delta_{u_j}}$ represents the mean value of the rating of u_i and u_j respectively.

2) the calculation of the user trust

In real life, when the user makes a choice, he may tend to accept the suggestion of the family, fiends, or the people with experience or authority. In online recommender system, the target user usually does not know most users. However, in some cases, it can infer whether a stranger is trustworthy according to the user interaction behavior or the user's own information.

Definition 2.1. [11]. *Trust. The quantification of the degree that u_i trusts u_j, and uses $T(u_i, u_j)$ to represent.*

The representation of the trust includes two categories: probabilistic method, the value of the trust is 1 or 0, that is a user is either trusted or untrusted; progressive method, it estimates the trust when the behavior can bring a certain positive effect, that is the information is right or wrong in a certain extent. And it uses different values to represent the different degrees of trust. For example, the method in [12] used four different values to represent the degree of trust: very trust, trust, distrust, very distrust.

In the paper, the two representation methods are combined to measure the user trust, and the value of the trust is set in $[0, 1]$. On the one hand, it calculates the user trust according to the social information of the user, and uses $t(u_i, u_j)$ to represent the obtained trust. Supposing u_i is a fan of u_j, but u_j is not a fan of u_i, when u_i is the target user, $t(u_i, u_j) = 1$, while when u_j is the target user, $t(u_j, u_i) = 0$, that is the trust is asymmetry.

On the other hand, the user's oneself trust is calculated according to the influence of the user in social network, and use $t'(u_j)$ to represent the obtained trust. In the paper, it chooses five basic features of the user: the number of fans, the number of reviews, the number of flowers, the value of contribution, the level of user community, and the formula is as follows:

$$t'(u_j) = \frac{f_{u_j}}{N_U} + \frac{d_{u_j}}{\sum\limits_{i=1}^{N_U} d_{u_i}} + \frac{h_{u_j}}{\sum\limits_{i=1}^{N_U} h_{u_i}} + \frac{g_{u_j}}{\sum\limits_{i=1}^{N_U} g_{u_i}} + \frac{m_{u_j}}{M} \tag{7}$$

where f_{u_j}, d_{u_j}, h_{u_j}, g_{u_j}, m_{u_j} represent the number of fans, the number of reviews, the number of flowers, the value of contribution, the level of user community; $N_U = |U|$ represents the number of all users; M represents the highest level of user community.

The value of the user trust $t'(u_j)$ obtained by Formula (7) may be greater than 1, which exceeds the setting range. Therefore, it needs to do normalization, and the formula is as follows:

$$t''(u_j) = \frac{t'(u_j) - t_{\min}}{t_{\max} - t_{\min}} \tag{8}$$

where $t_{\max}$ and $t_{\min}$ represent the maximum and minimum of the trust obtained by Formula (7).

The user trust $T(u_i, u_j)$ is calculated by fusing two obtained trust, and the formula is as follows:

$$T(u_i, u_j) = \frac{t(u_i, u_j) + t''(u_j)}{2} \tag{9}$$

The formula that is used to calculate user weight is as follows:

$$W(u_i, u_j) = \lambda_1 * Sim(u_i, u_j) + \lambda_2 * T(u_i, u_j) \tag{10}$$

where λ_1 and λ_2 represent the weight, and $\lambda_1 + \lambda_2 = 1$.

2.4. **The prediction of the user preference.** The prediction of user preference includes two step: firstly, it needs to find the nearest neighbors, and the user whose weight is greater than $0.2 * W(u_i, u_j)_{\max}$ is selected as the nearest neighbor of the target user, and $W(u_i, u_j)_{\max}$ is the maximum of the user weight; and then, it needs to predict the rating of the target user according to the ratings of the nearest neighbors.

According to the obtained user weight and the rating, the weighted average method is employed to predict the rating, and the formula is as follows:

$$\delta_{ij} = \overline{\delta_i} + \frac{\sum\limits_{u_k \in U'} W(u_i, u_k) * \left(\delta_{kj} - \overline{\delta_j}\right)}{\sum\limits_{u_k \in U'} W(u_i, u_k)} \tag{11}$$

where U' represents the set of the nearest neighbors of the target user; $\overline{\delta_i}$ represents the average of the normalized u_i's ratings. The predicted rating is as follows:

$$r_{ij} = round\left(\delta_{ij} * (r_{i,\max} - r_{i,\min}) + \overline{r_i}\right) \tag{12}$$

where $round()$ is the rounding function.

According to the predicted rating, it needs to rank the restaurant, and the top K restaurants are recommended to the target user.

3. Experiments and Analysis.

The data set is the real data of user and restaurant and is collected from dianping.com using the crawler software. The data set includes the background information (e.g., user ID, the number of fans, the number of flowers, the level of user community, the value of contribution, the tags of taste, average consumption), social information (e.g., following information, fans information) and the behavior information (the name of the rated restaurant, rating, rated time) in 6 months of 500 users. It also includes the basic information of 2000 restaurants: restaurant ID, location, type, average consumption. Due to considering the time decay, the leave-one method is employed to select the training set and test set. The data of the first 5 months is selected as the training set, and the data of the sixth month is selected as the test set.

3.1. **Evaluation.** F is employed to evaluate the effectiveness of the proposed method. The formula is as follows:

$$F = \frac{2 * P * R}{P + R} \tag{13}$$

where P represents the precision, R represents the recall, and the formulas are as follows:

$$P = \frac{N_a}{N_p} \tag{14}$$

$$R = \frac{N_a}{N_r} \tag{15}$$

where N_a represents the number of accurate predicted restaurants; N_p represents the number of recommended restaurants; N_r represents the number of the rated restaurants.

3.2. **Experimental step.** The experimental step is as follows. (1) The determination of θ. The role of θ is to reduce the impact of the similarity produced by the filled ratings, so the value of θ should be less than 1; when $\theta = 0$, it does not consider the filled rating, so the matrix of the user-restaurant is very sparsity. Hence, θ is set the value which is in $(0, 1)$, and the step is 0.1. In the step, the user weight only considers the impact of the user similarity, so $\lambda_1 = 1$; the number of the recommended restaurants K is set 5.

(2) The determination of λ_1 and λ_2. Due to $\lambda_2 = 1 - \lambda_1$, it is only necessary to set the value of λ_1. When $\lambda_1 = 0$, it does not consider the impact of the user similarity; when $\lambda_1 = 1$, it does not consider the impact of the user trust. Hence, similar with the

setting of θ, λ_1 is set the value which is in $(0, 1)$, and the step is 0.1; the number of the recommended restaurants K is set 5.

(3) The determination of K. When recommending the restaurant for the target user, if the number of recommended restaurants is very few, the recall is lower; if the number is very large, it loses the significance of the recommendation. Through analyzing the data set, the number that most users rate the restaurant usually is less than 5 per month. This is because the user usually goes to the restaurant with family or friends at weekend or holidays. Hence, the value of K should be greater than 5, but not too great. The value of K is set 5, 10, 15, 20.

(4) The comparison of different methods. The proposed method does not consider the context information, so the method in [4] and the traditional CF are selected as the comparative methods.

3.3. **Experimental results and analysis.** The experimental results are shown as Figures 1-4.

1) The impact of the parameter θ

From Figure 1, we can know that: when $\theta = 0.7$, the obtained results are the best. This is because when θ is relatively small, the impact of the filled rating is tiny. It mainly relies on the original rating when seeking the nearest neighbors of the target user, and the data is sparsity, so the accuracy of the obtained results is low; when $\theta > 0.7$, due to the error of the filled rating, it causes too many errors, so the accuracy of the obtained results gradually decrease.

2) The impact of the weight λ_1 and λ_2

Figure 2 shows that when $\lambda_1 = 0.6$, the obtained results are the best; when the value of λ_1 increases, the accuracy of the recommender system also increases; when $\lambda_1 > 0.6$, the accuracy begins to reduce. This is because when the value of λ_1 increases, the impact of the user trust gradually reduces, while the impact of the user similarity gradually increases. It shows that: (1) the user weight is not only impacted by the user similarity, but is also impacted by the user trust; (2) the impact of the user similarity is greater than that of user trust.

3) The impact of K

Figure 3 shows that: with the increasing of K, the recall gradually improves, while the precision gradually reduces; when $K = 10$, the value of F is the best, and improves 0.0035 than that when $K = 5$. This is because: the greater the value of K is, the more the

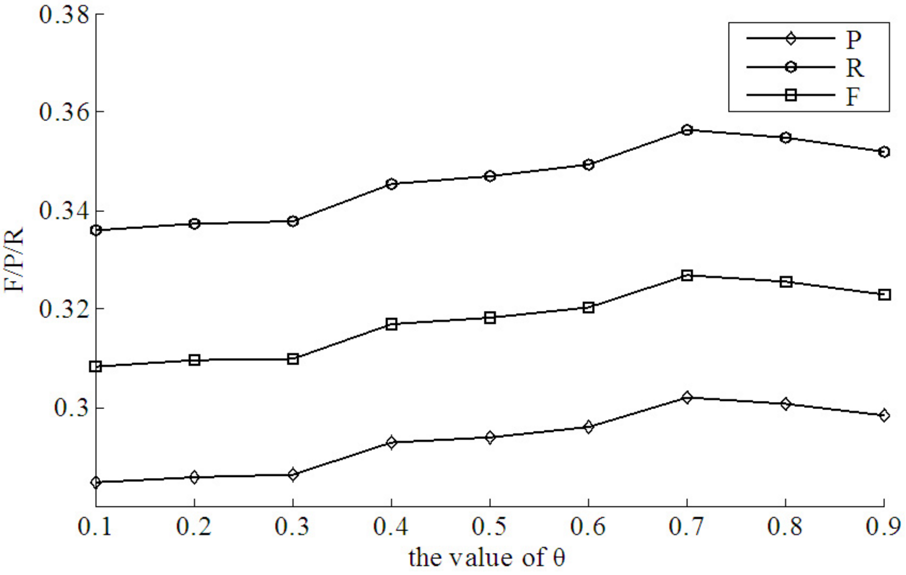

Figure 1. The results obtained when θ is set of different values

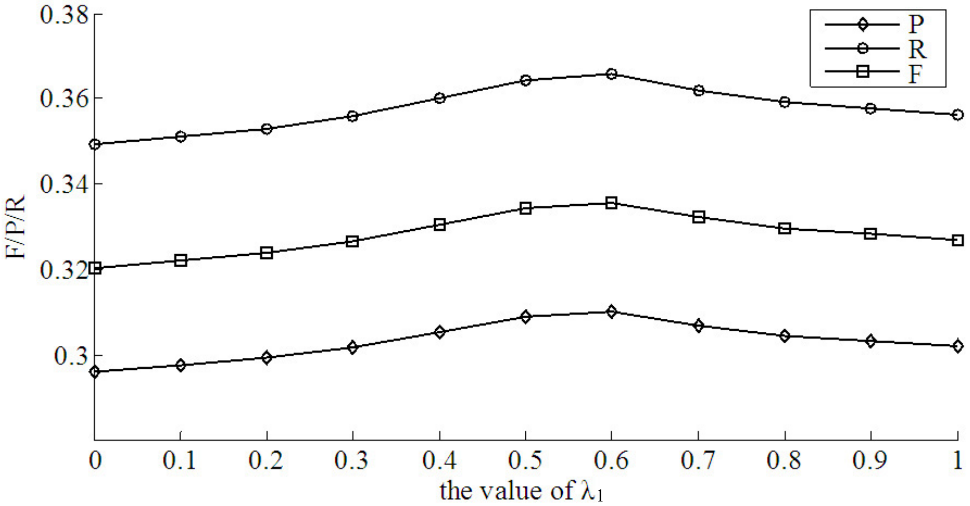

FIGURE 2. The results obtained when λ_1 is set of different values

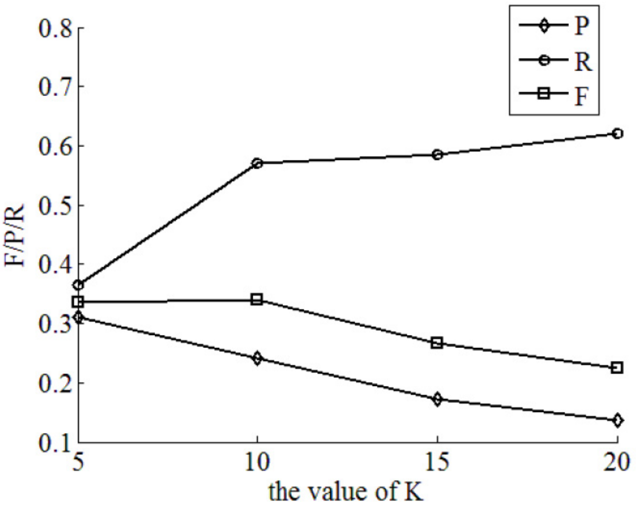

FIGURE 3. The results obtained when K is set of different values

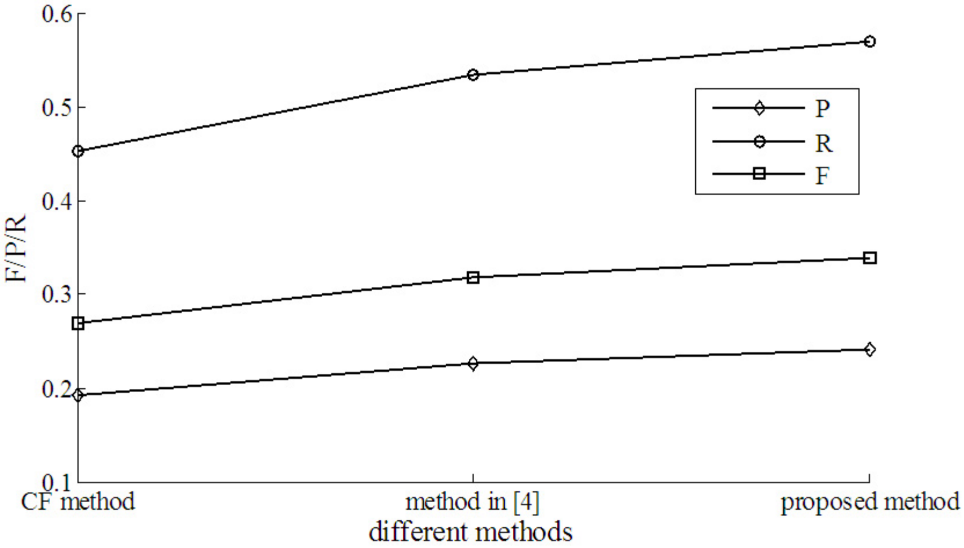

FIGURE 4. The obtained results by different methods

recommended restaurants are, so the number of accurately recommended restaurants also increases, while the number of the restaurant that the target user has rated is constant, hence, the recall will increase or keep constant with the increasing of the value of K. On the other hand, with the increase of the value of K, the number of recommended restaurants increases more than that of accurately recommended restaurants, so the precision decreases.

4) The comparison of different methods

Figure 4 shows that: compared with the results obtained by the traditional CF and the method in [4], the results obtained by the proposed method in the paper are the best; compared with these two methods, the value of F increases 0.0691 and 0.0211 respectively. This is because the traditional CF only considers the rating given by the user, and the data is relative sparsity, so the recommended results is the worst. The method in [4] considers the tags and the latent factor, so the obtained results are better than that obtained by the traditional CF. The proposed method in the paper not only considers the tags information of the user and the restaurant, but also considers the time and the user trust, so the obtained results are the best.

4. **Conclusions.** In order to recommend the personalized restaurant for user, a restaurant recommendation algorithm is proposed in the paper. In order to alleviate the sparsity problem, the improved null filling method is used to calculate the null rating according to the tags information; the trust of user oneself and the trust between users are considered when calculating the user trust; in order to accurately find the nearest neighbors of the target user, the similarity of user preference and the user trust are considered. The experimental results verify the proposed method is superior to the existing methods. According to the theoretical analysis and the experimental verification, we can obtain the conclusion: in the restaurant recommender system, it is necessary to consider the impact of time and user trust, and the user similarity plays a more important role than the user trust.

The proposed method does not consider the context and only considers the user trust simply. In the future work, it needs to research the propagation of the trust and the impact of context.

Acknowledgment. This work is partially by the National Natural Science Foundation of China (NSFC) (Grant No. 61402331, Grant No. 61402332, Grant No. 61272509) and the Key Technologies R&D Program of Tianjin (Grant No. 15ZCZDGX00200).

REFERENCES

[1] S. Park and J. Kang, Recommendation using analysis of semantic social network in social network services, *ICIC Express Letters*, vol.10, no.3, pp.547-553, 2016.

[2] P. Kantor, R. Francesco, R. Lior and B. Shapira, *Recommender Systems Handbook*, 2011.

[3] M. A. Hameed, O. A. Jadaan and S. Ramachandram, Collaborative filtering based recommendation system: A survey, *International Journal on Computer Science and Engineering*, vol.4, no.5, pp.859-876, 2012.

[4] M. Z. Ge, M. Elahi, I. F. Tobias, F. Ricci and D. Massimo, Using tags and latent factors in a food recommender system, *Proc. of the 5th International Conference on Digital Health*, Florence, Italy, pp.105-112, 2015.

[5] M. Elahi, M. Z. Ge, F. Ricci, I. F. Tobias, S. Berkovsky and M. David, Interaction design in a mobile food recommender system, *Proc. of the Joint Workshop on Interfaces and Human Decision Making for Recommender Systems*, Vienna, Austria, pp.49-52, 2015.

[6] W. T. Kuo, Y. L. Kuo, J. Y. Hsu and R. T. Tsai, Contextual restaurant recommendation utilizing implicit feedback, *Proc. of Wireless and Optical Communication Conference*, Taipei, Taiwan, pp.170-174, 2015.

[7] R. Katarya and O. Verma, Restaurant recommender system based on psychographic and demographic factors in mobile environment, *Proc. of the International Conference on Green Computing and Internet of Things*, Greater Noida, Delhi, India, pp.907-912, 2015.

[8] F. Z. Zhang, K. Zheng, N. J. Yuan, X. Xie, E. Chen and X. F. Zhou, A novelty-seeking based dining recommender system, *Proc. of the 24th International Conference on World Wide Web*, Florence, Italy, pp.1362-1372, 2015.

[9] Y. E. M. E. Alami, E. H. Nfaoui and O. E. Beqqali, Improving neighborhood-based collaborative filtering by a heuristic approach and an adjusted similarity measure, *Proc. of the International Conference on Big Data, Cloud and Applications*, Tetuan, Morocco, pp.25-26, 2015.

[10] G. Sun, *Research of Improved Recommendation Algorithm Based on Time Effect and Changes in Users Interest*, Beijing University of Posts and Telecommunications, Beijing, 2014.

[11] Z. B. Gan, C. Zeng, R. Ma and H. W. Lu, C2C e-commerce trust algorithm based on trust network, *Journal of Software*, vol.26, no.8, pp.1946-1959, 2015.

[12] M. Kim and P. Sang, Group affinity based social trust model for an intelligent movie recommender system, *Multimedia Tools & Applications*, vol.64, no.2, pp.505-516, 2013.