International Journal of Database Theory and Application
Vol.9, No.10 (2016), pp.75-86
http://dx.doi.org/10.14257/ijdta.2016.9.10.07

Research on a New Collaborative Filtering Recommendation Algorithm Based on Data Mining

Dong Liang

Qiongtai Teachers College, Haikou 570100, china,
18689851015@163.com

Abstract

Under the conditions of different community formation, this paper proposed two different models of formation communities. Firstly, we put forward two kinds of similarity calculation models, and compare them with the traditional similarity model, Secondly, several similarity models are tested under different conditions of community formation. Finally it compares tow models of forming communities and finds that for non-strict division of community model has a higher accuracy and diversity of recommendation, compared with the strict division of community model. Thus, the experiments show that the non-strictly divided communities' model is more suitable for recommendation system, especially for the personalized recommendation.

Keywords*: Recommendation system, Collaborative filtering, Data mining*

1. Introduction

The objective of individualized recommendation [1-5] is to boost sales on the basis of recommending items which can satisfy users' preference through collecting and analyze individual consumers' online behavior and purchasing record data, and withdraw their potential preferences. So far the common individual recommendation algorithm includes recommendation based on content and collaborative filtering recommendation. Content-based recommendation does to user items which match well with their preferences after acquiring their interests through recording their online browsing (like web pages they often click and what time they click) and purchasing logs [6-8]. Collaborative filtering algorithm, as its name implies, firstly analyzes user interests, then utilizes collaborative thinking to recognize users who have similar preference with target user or item similar to recommended one, next combines those similarity information to screen out items which are possibly interesting to users to finally complete prediction. GroupLens research team developed collaborative filtering system based on user rating. By advantage of its tremendous database information, the system is used to recommend movie and news, *e.g.* Douban, which recommends movie and music [9-10].

Although in research field, traditional recommendation method works wonderfully, it's not good to use for e-business platform, because it does better in acquiring user's rating of item and purchasing information [11-15]; other information is mostly hidden data like click, page detention time *etc.* With popularization of Web2.0, more and more websites allow users to give comprehensive rating of and comment on items after purchasing them. User review is information which reflects the most directly user's real preference; unfortunately at most websites, such kind of review is only limited to text review and holistic rating. Traditional recommendation algorithm considers only overall rating, neglecting lots of significant information latent in text review. Meanwhile, overall rating can hardly exhibit users' special preference for item's each feature. Hence individual recommendation can't be done well by merely depending user's overall rating [16-18].

However, with growing data scale, user data and the rapid enlargement of object data, collaborative filtering technology meets challenges. Due to huge matrix size, and user's participation information limited to a certain period, there would be the case that matrix

ISSN: 2005-4270 IJDTA

becomes more and more sparse. No matter what kind of similarity model is adopted, it's not possible to solve the problem of data sparsity; thus the recommendation effect degrades largely [19]. In this case, for enormous network data, especially the recommendation requirements based on Internet, a more efficient recommendation method is used rather than the content-based or collaborative filtering recommendation. Here we introduced the method based on community recommendation [20].

The core thinking of collaborative filtering recommendation system is to find out given user's similar (interest) user from user groups by analyzing user interest; then the system will have prediction of the user's degree of fondness of certain information through combining ratings of the information by those similar users. But with rapid development of Internet, traditional collaborative filtering recommendation system faces a skyrocketing number of users [21-22]. The collaborative filtering recommendation based on user interest mining not only needs to calculate the similarity among numerous users but also is challenged with online calculation of new coming users of a big quantity. Among existing recommendation systems, the commonest recommendation pattern is collaborative filtering recommendation based on item. The item-based collaborative filtering recommendation is capable to compute offline the similarity among rated items, which is helpful to enhance the response capability of the entire system.

Social network is defined as network based on interpersonal relationship, which includes individual association network and also that of small groups Community is the collection consisted of a few individuals. The construction of social network builds on the basis of collecting individual resources and information exchange [23]. If the model formed by community is employed for recommendation, it's likely to get together users with similar interests or features; then, provide recommendations to target users through user information in the community. That will reduce the complexity of calculation, improve working efficiency of the recommendation system, and contribute to in-depth exploration of relative information in the community [24-25].

2. Recommendation Algorithm Based on Community Relationship in Network

2.1. Improved Equation of User Similarity

1 Traditional similarity calculation (TSC)

Generally bipartite network includes user $U=\{u_1,u_2,...,u_i...,u_m\}$, object $O=\{o_1,o_2,...,o_p...,o_n\}$ and edges $E=\{e_{ij},:u\in U,o\in O\}$ joining them up. So in a bipartite network, the similarity between two users is calculated:

$$Sim_{TS}(u_i,u_j)=\frac{|C(u_i)\bigcap C(u_j)|}{|C(u_i)\bigcup C(u_j)|} \tag{1}$$

In the network, the number of movies chosen by each user is limited, because it's related with its time, energy, interest *etc.* If it's calculated with traditional equation of similarity, using denominator to divide the number of each selected movies, similarity becomes lower between users who watched more movies, but higher between users who watched fewer films. That is not logic. Considering shortcoming of traditional expression, we improved similarity calculation formula.

2 Improved similarity calculation formula (ISC)

User's rating of objects is mapped into a 2-point system. It's often found in a classical recommendation model. If object (*e.g.* movie) rating is five points, and scoring is reduced from a 5-point system to a 2-point system where there's only 0 and 1, it means only need

to consider whether user loves the object. The point not less than 3 suggests that user likes the object; otherwise, user dislikes it.

Although the method can reduce computer processing speed and increase running efficiency of the recommendation system, in order to enhance the accuracy of calculating similarity, comprehensive rating information should be used instead of condensed information which cannot be complete. Hence, we present an improved formula to calculate the similarity. It makes full use of user's all rating information.

To compute similarity between two users, we can estimate differentiation between them. The difference between user u_i and u_j is defined as follows:

$$D_{IS}(u_i,u_j)=\frac{\sum_{k}^{|C(u_i)\bigcap C(u_j)|}|R_{i,k}-R_{j,k}|}{|C(u_i)\bigcap C(u_j)|.(R_{M\,\mathrm{ax}}-R_{Min})} \tag{2}$$

3 Similarity calculation formula with fault-tolerant rating (IST)

To the improved similarity calculation method, an approach with fault-tolerant mechanism is introduced. Considering that each user's evaluation may be arbitrary and faulty, for instance, when a user loves a movie but not very much, the rating is usually 3 or 4 points. A movie rated by 4 points may be not better than one by 3 points; or a movie rated by 3 points may not be worse than one by 4 points. In this case, we bring $\Delta R=1$ as fault-tolerant rating. The difference degree between user u_i and u_j is defined as follows:

$$D_{IST}(u_i,u_j)=\frac{\sum_{k}^{|C(u_i)\bigcap C(u_j)|}|R_{i,k}-R_{j,k}\pm\Delta R|}{|C(u_i)\bigcap C(u_j)|.(R_{M\,\mathrm{ax}}-R_{Min})} \tag{3}$$

To classify user accurately to the belonged community, it's necessary to consider its connection with the community and define the similarity degree between user and community and that among communities.

(1) Similarity between user and community

With existing formula for calculating similarity between users, we can get the similarity degree between any user and one community, by the expression:

$$Sim_{UC}(u_i,C_8)=\sum_{u_k\in C_8}\frac{Sim(u_i,u_k)}{\|C_g\|} \tag{4}$$

By calculating the average value of the similarity between the users and the community, to determine the degree of association.

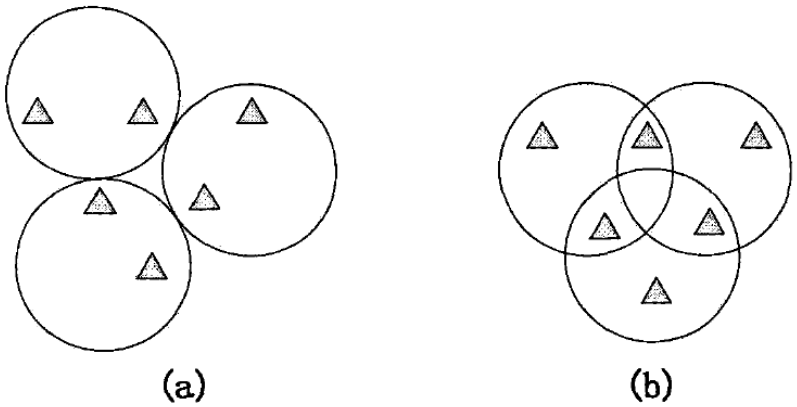
(2) Similarity among communities

$$Sim_{CC}(C_g,C_h)=\sum_{u_i\in C_g,u_j\in C_h}\frac{Sim(u_i,u_j)}{\|C_g\|.\|C_h\|} \tag{5}$$

The formula calculates the similarity between two groups of users, to determine the degree of correlation between the two groups.

2.2. Forming Process of Community

We utilized here two kinds of community models: strictly divided and not strictly divided community mode. In the former model, the membership degree of users in the community is 1, meaning one user can belong to only one community; in the later model, the membership degree of users is bigger or equal to 1, meaning one user can belong to several different communities at the same time. It is shown in Figure1.



(a) The Non-Strict Division of the Community (B) Comparison Chart

Figure 1. Strict Division of the Community

1. Formation of strictly divided community

(a)At initial stage, each single one is regarded as a community; if assume m units waiting for classification, then there're m independent communities, *i.e*. the number of community is initially m;

(b)Compute similarity among communities, which have the three cases:

(i) similarity between two independent units;

(ii) similarity between single unit and community with formula 1-2;

(iii) similarity between community and community with formula 4-5;

(c)From similarity SIM matrix got in the above, get element with biggest number; then mix together two communities of which the line and column that element belongs to to form a new community; at this moment, the total number of community lessens 1 and dimensions of similarity matrix reduces 1;

(d)Start from b) to repeat computation till convergence condition is sufficed.

Discussion of convergence condition

Suppose in the process the number of initial individuals is m; by now there're totally m communities. Whenever two communities are fused to constitute a new one, the totality of community cuts 1 down. Without limitation of convergence condition, the process will go on till all individuals are merged into a community.

2 Forming process of not strictly divided community

(a) At initial stage, each individual looked as an independent community; if there're m individuals, then there're totally m communities;

(b) Calculate similarity among communities; according to the similarity calculation formula, we can get similarity matrix;

(c) From similarity matrix $SimB_k$, individuals represented by elements which have bigger value than threshold are selected and put into the most similar community.

3 Generate recommendation list

We take one individual (user) x for instance. By referring to objects chosen jointly by other members in its community, we accumulate how many times such an object is chosen; then based on that, we recommend the first L objects with biggest number as its recommendation list. Formula is as follows.

$$rec(x) = \arg\max \left\{ \sum_{y} R_{y,k}, y : x, y \in C(x), x \neq y \right\} \tag{6}$$

3. Experiment Design and Discussion

3.1 Experimental Dataset

The dataset for the experiment was collected from MOVIELENS, which includes ratings of 1590 movies by 890 users. The rating quantity is over 200000. From that we chose randomly 10 groups of data, in the following conditions, it is shown in Table1.

User quantity reaches 200 in each group;

In each group, each user rates at least over 30 movies;

In each group, each user rates at least 2000 movies;

In each group, the sparsity of network constituted by user and movie is not lower than 5%.

Table 1. Statistics of the Data in Each Group

No.	Users	Objects	Links	Sparsity(%)
1	200	2267	9666	7.53
2	200	2207	9544	7.90
3	200	2256	8455	7.15
4	200	2312	9900	7.99
5	200	2145	7900	6.78
6	200	2356	8455	6.98
7	200	2767	7905	6.84
8	200	2215	9056	7.37
9	200	2345	9561	6.90
10	200	2289	8566	76.88

Then each group of data is divided into training data set (80%) and test data set(20%) according to a certain proportion. The recommended list of the system used is L=10.

3.2 Evaluation of the Effect of Recommendation

1 Precision

The Precision rate of the recommendation system is in Formula7:

$$P = \frac{1}{m} \cdot \frac{\sum_{r}^{m} d_r}{L} \tag{7}$$

2 Diversity of recommendation

In view of features based on community recommendation, we use average intra-user diversity as the measuring method of system recommendation result. The similarity formula between two objects is defined as follows:

$$Sim_{Diversity}(o_p, o_q) = \sum_{u=1}^{m} \frac{a_{u,p} . a_{u,q}}{\sqrt{k(o_p).k(o_q)}} \tag{8}$$

In the algorithm based on community recommendation, since it's not possible to ensure that enough long recommendation list is provided to each user, so the length of such list is $L_u^{'}$, referring to the length of recommendation list to user u. And each user in the system acquires different long recommendation list, which depends on whether the number of selected object to which each user belongs is bigger than L. Now we can get the diversity measuring of recommendation system result.

$$D_{Diverisity} = \arg\{D_{Diverisity}(u)\} = \frac{1}{m^{'}} \cdot \sum_{u=1}^{m^{'}} D_{Diverisity}(u) \tag{9}$$

3.3. Recommendation Result Test

1 Strictly divided community

Variation curve of recommendation precision plotted with dataset (Table1), it is shown in Figure2.

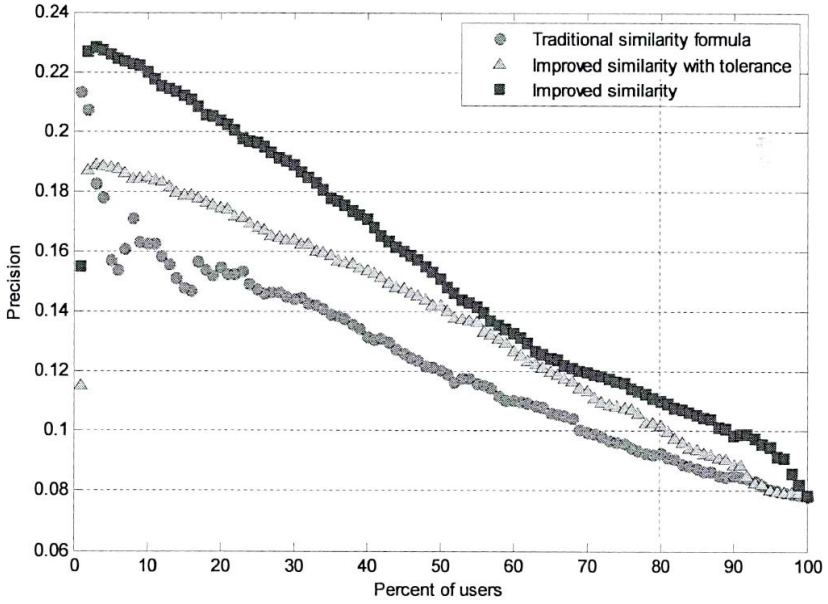


Figure 2. Comparison of Three Kinds of Recommendation Models Based on Strict Division of Community

There are three models based on traditional similarity, improved similarity and the one with fault-tolerant rating. We need to show accuracy variation of three models in one graph based on the strictly divided community. In it, the abscissa is percentage of acquired recommendation users based on community, *i.e.* threshold value of recommendation convergence. Take x-coordinate 80 for instance. It shows the stop condition of community aggregation process is: 80% users get the ability to be recommended by system. The vertical coordinate stands for relevant recommendation accuracy.

The improved model and the one with fault-tolerant mechanism achieve higher recommendation accuracy rate than the traditional model. Of that, the fault-tolerant mechanism enhances similarity degree between any two users due to introducing ΔR, causing that two similar users become too close and avoiding irrelevant users from getting far away. Hence its recommendation accuracy becomes lower than the improved one.

For the threshold value, if the 80% threshold, the user can be said to be recommended for 80%. If you have 85% or 90% as a threshold, the recommended range of users is expanded. But the accuracy of the recommendation has dropped. If 80% as the threshold, this time not only to meet the most recommended needs, and meet certain recommendation accuracy, as shown in the following Table2:

Table 2. Precision Comparison of Different Recommendation Algorithm

	Traditional similarity	Improved model	The similarity of fault tolerance mechanism	Heat conduction model	Probability transfer model
Precision	0.0824	0.0112	0.0104	0.0009	0.0116

A variety of curves based on the data set (Table1), it is shown in Figure3.

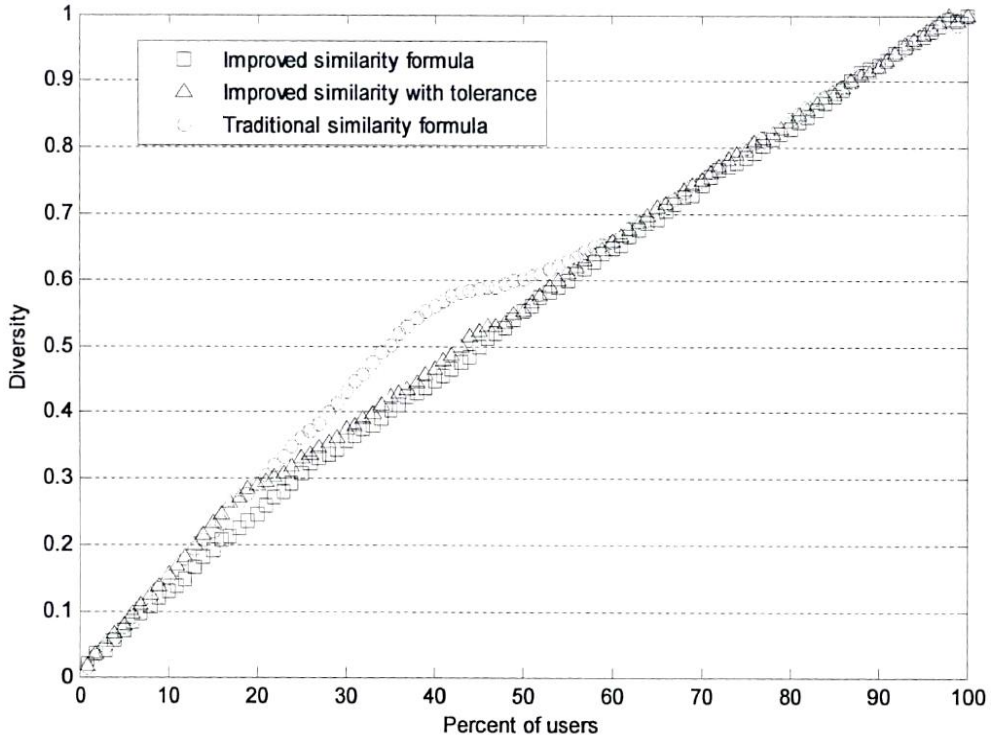


Figure 3. Comparison of Three Kinds of Recommendation Models Based on Strict Division of Community

In Figure3. Three similarity models show identical tendencies: with increasing percentage of recommendable users, the diversity of recommendation system becomes greater. At initial stage, when the threshold is 0%, *i.e*. the fewest users are recommended, the three models mentioned above all have initial recommendation diversity of 0.0148, 0.0158, 0.0198. At the ending stage, when threshold is 100%, meaning all users can be recommended. At this moment, three models have the same diversity of recommendation, which all is 0.99, meaning the object collection chosen by most individuals in the community used as content of recommendation list. By now the diversity of three models is of the same and approximates 1.

2 Not-strictly divided community

With ten groups of data in dataset (Table1), the accuracy changing curve of three different similarity models is portrayed in the condition of not strictly divided community. It is shown in Figure4.

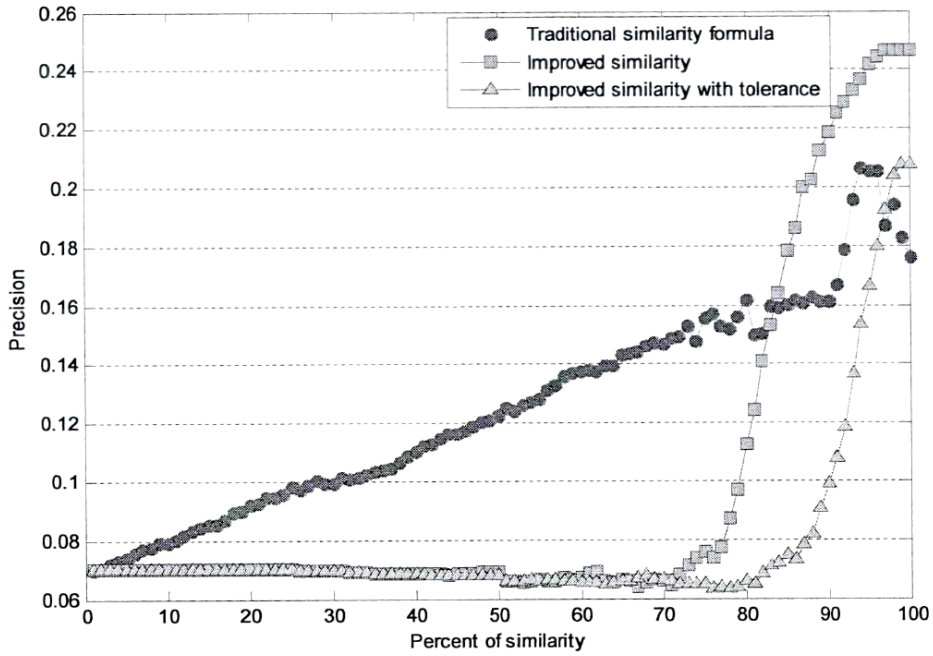


Figure 4. Precision Change of Three Similarity Models under the Non-Strict Division of Community

Figure 4 displays its x-coordinate differs from its previous horizontal coordinate in Figure2. The previous percentage of recommended users is abscissa; while in the picture the percentage of similarity is abscissa, as explained as follows. Since this bases on non-strict community, the similarity threshold between users in every community in that condition is considered as the judgment basis of community convergence. The calculating method of threshold is: using the differentials between the maximum similarity value and the minimum as change interval, with one hundred percentage of such interval as step length. Formula is as follows

$$\Delta(C_g) = \frac{Max(C_g) - Min(C_g)}{100} \times 100\% \tag{10}$$

For the three similarity models, they have the same recommendation accuracy rate at the initial phase, which is 0.7, because in the beginning, no matter which model is adopted, each community contains the same many member individuals. That's why the recommendation result is no difference.

With the increase of the horizontal coordinates, the accuracy of the recommendation increases gradually. If the 80% threshold, this can not only meet the most recommended requirements, and meet the recommendation precision. The precision of the recommendation is compared with other models, it is shown in Table3.

Table 3. Precision Comparison of Different Recommendation Algorithm

	Traditional similarity	Improved model	The similarity of fault tolerance mechanism	Heat conduction model	Probability transfer model
Precision ($P^{'}$)	0.1709	0.0114	0.0064	0.0086	0.0105

The variation curves of the recommended system are shown in Figure5.

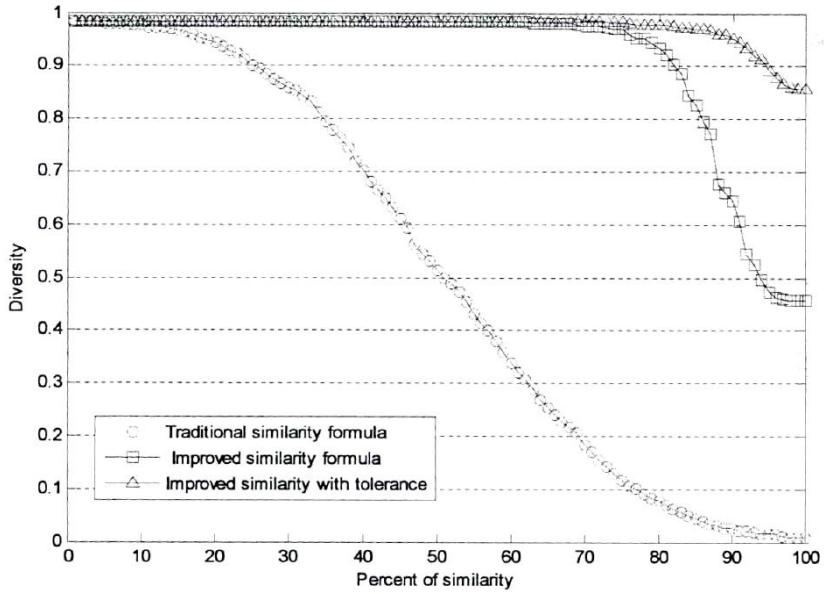


Figure 5. The Three Kinds of Similarity Model of the Recommended Diversity of the Curve Chart Based on the Non-Strict Division of the Community Conditions

It can see in Figure5, the proposed model's diversity value is lower than the other two. And with growing percentage of similarity degree at x-axis coordinate, the diversity value tends to be lower and lower. Initially, three models share common value of diversity, close to 1, indicating that the recommendation was gained, however, the individualization was still low. If we keep on regarding 80% of numerical values as percentage threshold of similarity, then it's possibility to obtain the diversity situation of those similarity models and compare with known recommendation models. It is shown in Table4.

Table 4. The Degree of Diversity of Recommendation System Under Different Recommendation Model

	Traditional similarity	Improved model	The similarity of fault tolerance mechanism	Heat conduction model	Probability transfer model
Strict division of community	0.845	0.818	0.829	0.809	0.715
Non-strict division of community	0.079	0.945	0.984	0.809	0.715

4. Conclusion

In this paper, two kinds of different community formation models are proposed, and the application and recommendation of the three models are compared with the two models. By using the data of the MOVIELENS data set, it is verified that the model based on the community formation is not only in the recommendation accuracy.

Which model is more suitable for the system to do the system recommendation: for the strict division of the community model, although the initial stage of the initial stage of the aggregation of the recommendation is very high,

But at the same time, the number of users can be recommended is too sparse, it cannot meet the actual needs of the recommendation;

References

[1] N. Z. Jiafeng and T. Strong, "Based on the emotion lexis of online product reviews personalized recommendation method", Zhengzhou University Journal (NATURAL SCIENCE EDITION), vol. 2, **(2011)**, pp. 48-51.
[2] N. Z. Jiafeng, "Online consumer reviews", Fuzzy intelligent product recommendation system based on system engineering, vol. 11, **(2013)**, pp. 116-120.
[3] T. Xueqing and H. Shan, "A review of the research on music personalized recommendation system", modern library and information technology, vol. 9, **(2014)**, pp. 22-32.
[4] W. Wei, W. Hongwei and M. Yuan, "Collaborative filtering recommendation algorithm research: Considering online review of affective tendency", System engineering theory and practice, vol. 12, **(2014)**, pp. 3238-3249.
[5] H. H. Yan, "Visual analysis of online social network", Journal of the Chinese Academy of Sciences, vol. 2, **(2015)**, pp. 229-237.
[6] Y. Ming, Q. Wei, Y. Xiangbin and L. Yijun, "Journal of Management Sciences in China analysis of online product reviews the utility", vol. 5, **(2012)**, pp. 65-75.
[7] Y. Li, "Research on Personalized Recommendation Algorithm Based on clustering of collaborative filtering", Huazhong Normal University, **(2014)**.
[8] Z. Liang, "Research on the recommendation algorithm based on collaborative filtering and clustering", Jilin University, **(2014)**.
[9] Z. Jing, "Research on Recommendation Algorithm of collaborative filtering based on trust", Yanshan University, **(2013)**.
[10] L. Xueqin, "Research on Personalized Recommendation Algorithm Based on location service", South China University of Technology, **(2012)**.
[11] Y. Jing, "User interest model and real-time personalized recommendation algorithm research", Nanjing University of Posts and Telecommunications, **(2013)**.
[12] W. Wenhao, "Research and application of personalized recommendation algorithm based on mobile library", China University of Geosciences (Beijing), **(2015)**.
[13] W. Cong, "Recommended. Research on Algorithm of personalized user behavior", Harbin Institute of Technology, **(2015)**.
[14] F. Pengcheng, "Research on Personalized Recommendation Algorithm Based on context aware", Donghua University, **(2014)**.
[15] Z. Lan, "Research on Personalized Recommendation Algorithm Based on collaborative filtering", Huazhong Normal University, **(2009)**.
[16] L. Qingwen, "Research on the recommendation algorithm based on collaborative filtering", University of Science & Technology China, **(2013)**.
[17] S. Guangfu, W. Yue, L. Qi, Z. Chen and C. Enhong, "A collaborative filtering recommendation algorithm based on sequential behavior", Journal of software, vol. 11, **(2013)**, pp. 2721-2733.
[18] Sui Geng. Research on E-commerce Recommendation Algorithm Based on collaborative filtering. Shandong Normal University, **(2014)**.
[19] H. Yang, "Research on Collaborative Filtering Recommendation Algorithm Based on item clustering and preference categories", Zhejiang Sci-Tech University, **(2014)**.
[20] J. Herlocker, J. Konstan and L. Terveen, "Evaluating Collaborative Filtering Recommender System", ACM Trans on Information System (TOIS), vol. 22, no. 1, **(2004)**, pp. 5-53
[21] R. Burke, "Hybird System for Personalized Recommendations. Intelligent Techniques for Web Personalization", Springer Berlin Heidelberg, **(2005)**, pp. 133-152
[22] Huang Yongwen. Research on Key Technologies of Chinese product reviews mining. Chongqing: Chongqing University, 2009.
[23] T. Jun and Z. Ning, "A collaborative filtering recommendation algorithm based on user interest classification", Computer system application, vol. 5, **(2011)**, pp. 55-59.
[24] W. Xinjun, W. Hongchen, C. Yong and P. Zhaohui, "Improved recommendation algorithm based on collaborative filtering and clustering", Computer research and development, vol. 3, **(2011)**, p. 205-212.
[25] W. Yubin, M. Xiangwu and H. Hoon, "A information aging based collaborative filtering recommendation algorithm", Journal of electronics and information technology, vol. 10, **(2013)**, pp. 2391-2396.

Author

Dong Liang, He received his B.S degree from Hunan Institute of Science and Technology and received his M.S degree from Chongqing University. He is a lecturer at Qiongtai Teachers College. His research interests include computer application.