International Journal of Smart Business and Technology
Vol.1, No.1 (2013), pp.39-48
http://dx.doi.org/10.21742/ijsbt.2013.1.1.04

User Preference Collaborative Filtering Recommendation Algorithm based on Data Mining

Andrew M. Barthelemy[1] and George Suter[2]

[1,2]University of Adelaide, Adelaide, Australia
[1]andrbarthelemy371@gmail.com, [2]george.suter@adelaide.edu.au

Abstract

With the rapid development of information technology, the Internet has developed into the most important e-commerce platform. This article integrates user preference mining technology into collaborative filtering recommendations and proposes an e-commerce collaborative filtering recommendation algorithm based on user preference mining. This algorithm Aiming at the traditional collaborative filtering recommendation algorithm that only uses the user's explicit preference information when calculating user similarity, and ignores the user's implicit preference knowledge, it is proposed to use user preference mining technology to perform user explicit preference knowledge and implicit preference information. Mining preference knowledge, using the excavated user preference knowledge to calculate user similarity, and realizing the nearest neighbor community formation mechanism based on user preference knowledge. On this basis, intelligent recommendation of user needs is realized. Experiments show that the algorithm has achieved expectations Effective, comprehensive use of user preference knowledge for collaborative filtering recommendation is the key to improving the accuracy and quality of recommendation results.

Keywords: *E-commerce, Collaborative filtering, Preference mining, Recommendation*

1. Introduction

With the acceleration of the popularization of the Internet and the rapid development of information technology, the Internet has developed into the most important e-commerce platform. With the expansion of the network scale and the increase of online shopping users, how to use it in the numerous e-commerce websites and massive commodities Quickly and accurately select products that meet user needs, and automatically and intelligently recommend them to target users, which has become an urgent problem in the current global network economy and e-commerce development.

E-commerce collaborative filtering recommendation is the most important one of the current e-commerce recommendation methods [1]. Compared with content-based recommendation methods and rule-based recommendation methods, this method can not only recommend the same or similar to users according to their preferences. Commodities can also recommend products of interest to users. Since the first proposed by Goldberg et al. [2] in 1992, collaborative filtering recommendation methods have always been a research hotspot in the field of e-commerce recommendation [3], and its research content mainly focuses on three aspects:

Article history:
Received (July 16, 2013), Review Result (September 21, 2013), Accepted (November 1, 2013)

Print ISSN: 2288-8969, eISSN: 2207-516X IJSBT

(1) User similarity calculation in the collaborative filtering recommendation process. For example, Sarwar et al. [4] used the SVD (Singular Value Decomposition, SVD) method to reduce the dimensionality of the features of user rating data and obtain the latent semantics between different users. On this basis, the user similarity calculation is performed, and the processing efficiency and accuracy of the collaborative filtering recommendation system are improved through this processing. Ren at al. [5] proposed user similarity calculation based on knowledge processing mechanism, using cloud model for qualitative knowledge representation and qualitative and quantitative knowledge conversion, solving the shortcomings of traditional user similarity calculation methods, and realizing collaborative filtering recommendation based on cloud model. Massa et al. [6] The proposed tag-based collaborative recommendation algorithm for the public tagging system uses the extended PLSA model to map user tags to semantic topics with clear semantics, eliminates the semantic ambiguity of tags, and realizes user similarity calculation at the semantic level.

(2) User trust calculation in the process of collaborative filtering recommendation. For example, the trust-based collaborative filtering recommendation method proposed by Massa et al. [6] provides recommendation services based on the trust relationship between users and improves the quality and efficiency of recommendation services. Kazienko et al. [8] The proposed method of e-commerce diversity recommendation based on trust, realizes the diversity of collaborative recommendation by selecting trusted neighbors with good subject diversity. The collaborative recommendation method based on trust propagation proposed by Madani et al. [9] compares the idea of trust propagation with PageRank. In combination, by recommending potential friends who may be of interest to the user, the user's degree of interaction and user stickiness is improved.

(3) User preference calculation in the process of collaborative filtering recommendation. For example, Kim et al. [10] proposed an optimized collaborative recommendation algorithm based on user interest and characteristics, which calculates the user's interest in product items to group users and uses Bayesian algorithm analysis The degree of user preference for products with different characteristics. Strunjas et al. [11] proposed a collaborative filtering method based on multi-objective optimization bi-clustering, through the analysis of user registration information, to obtain user preferences and needs, according to these preferences and needs for targeted Project recommendation. Wu at al. [12] proposed a collaborative filtering recommendation algorithm based on comprehensive interest, which integrates the explicit and invisible preferences of users for accurate positioning and efficient recommendation of target users.

Since the collaborative filtering recommendation method based on user preferences can automatically provide users with the commodities or resources they need according to their preferences or needs, it provides technical support for solving the problem of information overload on the Internet. However, the current recommendation system uses Most of the user preference information is the user's explicit preferences, such as user ratings and voting information, while the user's invisible preference information, such as page stay time, purchase motivation, and other information is not taken into consideration. At the same time, with the passage of time and the changes in user awareness, user preferences will also change accordingly. This requires the recommendation system to be able to predict user preferences based on the user's change information. Through in-depth mining and analysis of user preference information, especially invisible preference information and preference dynamic change information, grasping user preferences and needs on time is the key to improving the service quality and efficiency of the recommendation system.

2. E-commerce collaborative filtering recommendation algorithm based on user preference mining

E-commerce Collaborative Filtering Recommendation Algorithm Based on User Preference Mining (ECFR-UPM), whose basic idea is to use data mining and other technologies to conduct in-depth mining and analysis of user preference information to obtain users Preference knowledge. Use the acquired user preference knowledge to construct the user preference space matrix to calculate the user similarity and obtain the nearest neighbor communities with the same or similar preferences. Predict the needs of target users based on the buying behavior and preference knowledge of the nearest neighbor community and realize automated intelligent recommendations.

2.1. User preference mining

User preference mining refers to the use of data mining, behavior analysis, trend prediction, and other technologies to conduct in-depth mining and analysis of user preference information to obtain user preference knowledge. Traditional user preference mining mainly uses explicit information such as user registration information, user ratings, user reviews, etc. The analysis and processing of the user's explicit preference knowledge are not enough. The analysis and processing of the user's stay time on the Web page, the number of clicks, and other hidden information to obtain the user's implicit preference knowledge are not enough. The mining technology mines the user's explicit preference information and obtains the user's explicit preference knowledge. Use behavior analysis and Weblog mining technology to mine the user's implicit preference information and obtain the user's implicit preference knowledge. Use trend prediction and association mining technology to mine trend information such as changes in user browsing behavior and unrated items, obtain knowledge of user preference changes, and reasonably predict user preferences. The processing flow is shown in [Figure 1].

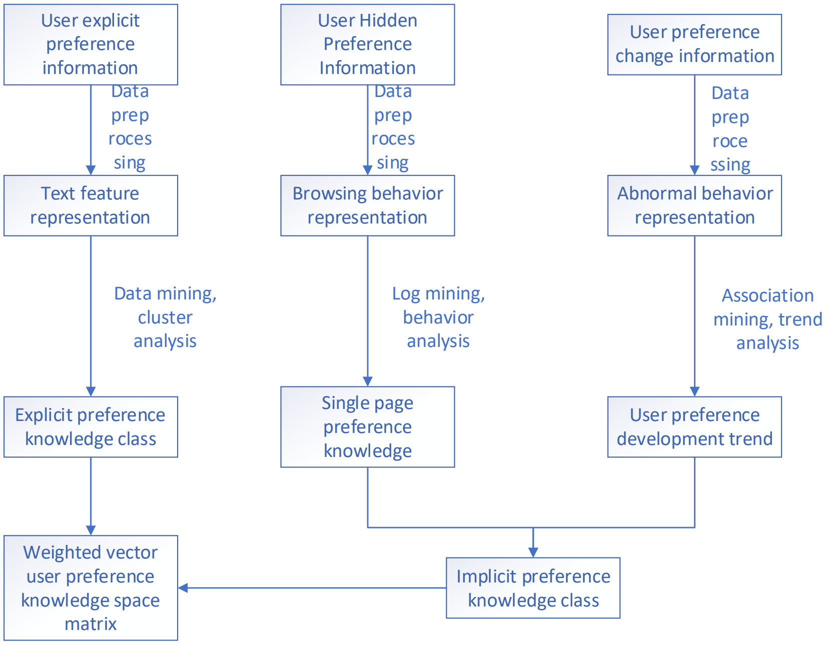
User explicit preference information
Data prep roces sing
Text feature representation
Data mining, cluster analysis
Explicit preference knowledge class
Weighted vector user preference knowledge space matrix
User Hidden Preference Information
Data prep roces sing
Browsing behavior representation
Log mining, behavior analysis
Single page preference knowledge
User preference change information
Data prep roce ssing
Abnormal behavior representation
Association mining, trend analysis
User preference development trend
Implicit preference knowledge class

Figure 1. Flow chart of user preference mining

(1) Explicit preference knowledge mining. The main object of explicit preference knowledge mining is explicit preference information such as the text information of the pages browsed by users, published comments, etc. This paper uses the K-means clustering algorithm to mine and analyze user explicit preference information. Acquire user preference knowledge clusters. Suppose the collected user explicit preference information is preprocessed to form a text document set $\mathrm{D} = \{\mathrm{d}_1, \mathrm{d}_2, \cdots, \mathrm{d}_\mathrm{n}\}$, then the process of user explicit preference knowledge mining is: First, construct the initial cluster $\mathrm{C} = \{\mathrm{c}_1, \mathrm{c}_2, \cdots, \mathrm{c}_\mathrm{i}, \cdots, \mathrm{c}_\mathrm{n}\}$, $\mathrm{c}_\mathrm{i} = \{\mathrm{d}_\mathrm{i}\}$, that is, treat all documents in D as a single initial user preference category; secondly, calculate the distance between any categories The similarity of $\mathrm{sim}(c_i, c_j)$, the initial clusters are merged and optimized according to the set threshold ε, that is, the category with the greatest similarity is selected $\max = \mathrm{Max}\{\mathrm{sim}(c_i, c_j)\}$, if $\max > \varepsilon$, then merge c_i and c_j into a new category $\mathrm{c}_\mathrm{k} = \mathrm{c}_\mathrm{i} \cup \mathrm{c}_\mathrm{j}$; if $\max < \varepsilon$, keep the categories of c_i and c_j, and finally divide D into a new sub-category cluster $C = \{c_1, c_2, \cdots, c_k\}$; Finally, take the subcategory C obtained above as the seed set $S = \{s_1, s_2, \cdots, \mathrm{s}_\mathrm{k}\}$ of the initial clustering center of the K-means algorithm, and calculate the similarity of $\mathrm{sim}(\mathrm{d}_\mathrm{i}, \mathrm{s}_\mathrm{i})$ between d_i and s_i, select the seed with the greatest similarity $\max = \mathrm{sim}(\mathrm{d}_\mathrm{i}, \mathrm{s}_\mathrm{i})$, assign d_i to the category c_i with s_i as the clustering center, and get the final clustering result $\mathrm{C}^* = \{\mathrm{c}_1, \mathrm{c}_2, \cdots, \mathrm{c}_\mathrm{k}\}$.

(2) Tacit preference knowledge mining. Many actions of users when browsing e-commerce websites can imply user preferences, such as staying time on related Web pages, number of visits, clicks, etc. [13]. This article uses similar behavior sequence clusters The comprehensive similarity calculation of user implicit preferences is carried out in two aspects: degree and similarity between clusters and the sequence aggregation algorithm of K-center aggregation are used to mine user groups with the same implicit preference and user preference knowledge of a single page. Reference [14] The data preprocessing method for network social user interest mining preprocesses the collected user implicit preference information, and obtains the user behavior sequence set $\mathrm{D} = \{\mathrm{d}_1, \mathrm{d}_2, \cdots, \mathrm{d}_\mathrm{n}\}$, the initial number of clusters is k, and the calculated behavior clusters Internal similarity:

$$S_w(k_i) = \frac{\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \{Sim(d_i, d_j)\}}{C_n^2} \tag{1}$$

Where, k_i represents the i-th cluster in the clustering mode when the number of clusters is k, and $S_w(k_i)$ represents the average similarity within the i-th cluster. The calculation of similarity between clusters uses a similar calculation method as that within clusters:

$$S_b(k_i, k_j) = \frac{\sum_{s=1}^{k_i} \sum_{t=1}^{k_j} \{Sim(d_i, d_j)\}}{d_i * d_j} \tag{2}$$

Where, $S_b(k_i, k_j)$ represents the similarity between the behavior sequence cluster i and family j. Through the mining and analysis of user behavior data, not only can the user's potential preferences and needs be obtained, but also based on the user's browsing time and clicks Abnormal information such as quantity predicts the changes of user preferences, and at the same time, it is possible to predict the user's explicit preferences and behavioral analysis for product items that users have not evaluated, to obtain the user's implicit preference knowledge.

After the above processing, the weighted keyword vector model is used to construct the user preference space matrix:

$$UPS = \{(i_1, w_1), (i_2, w_2), \ldots, (i_k, w_k)\} \tag{3}$$

Where, i_k represents the k-th preference type of the user, and w_k represents the weight of the k-th preference type in the user's preference.

2.2. Formation of nearest neighbor communities

The key to e-commerce collaborative filtering recommendation is to accurately locate the nearest neighbor of the target user. The basis for determining the nearest neighbor is to calculate the similarity between users. There are three main calculation methods commonly used:

(1) Pearson correlation similarity. The calculation formula is:

$$sim(u,v)=\frac{\sum_{\alpha\in P_w}\left(R_{u,\alpha}-\overleftarrow{R}_u\right)\left(R_{v,\alpha}-\overleftarrow{R}_v\right)}{\sqrt{\sum_{\alpha\in P_w}\left(R_{u,\alpha}-\overleftarrow{R}_u\right)^2}\sqrt{\sum_{\alpha\in P_w}\left(R_{v,\alpha}-\overleftarrow{R}_v\right)^2}} \tag{4}$$

Where, P_{uv} represents the set of scoring items, $R_{u,\alpha}$ and $R_{v,\alpha}$ represents the score of the product item α, and $\overleftarrow{R}$ represents the average score.

(2) Cosine similarity. Assuming that the scores of users u and v on the n-dimensional project space are m、n, the similarity between users u and v is:

$$sim(u,v)=\frac{m*n}{\parallel m\parallel*\parallel n\parallel} \tag{5}$$

(3) Modified cosine similarity. The modified cosine similarity fully considers the rating methods of different users, and its calculation formula is:

$$sim(u,v)=\frac{\sum_{\alpha\in P_w}\left(R_{u,\alpha}-\overleftarrow{R}_u\right)\left(R_{v,\alpha}-\overleftarrow{R}_v\right)}{\sqrt{\sum_{\alpha\in P_u}\left(R_{u,\alpha}-\overleftarrow{R}_u\right)^2}\sqrt{\sum_{\alpha\in P_v}\left(R_{v,\alpha}-\overleftarrow{R}_v\right)^2}} \tag{6}$$

Where, P_{uv} represents the set of scoring items, P_u and P_v represent the set of items rated by users u and v, respectively, $R_{u,\alpha}$ and $R_{v,\alpha}$ represent the rating of the product item α, and $\overleftarrow{R}$ represents the average rating.

Based on the modified cosine similarity, this paper incorporates user preference knowledge into the calculation of user similarity. The calculation formula is:

$$sim(u,v)=\frac{\sum_{i\in UPS_w}\left(w_{u,i}-\overleftarrow{w}_u\right)\left(w_{v,i}-\overleftarrow{w}_v\right)}{\sqrt{\sum_{i\in UPS_u}\left(w_{u,i}-\overleftarrow{w}_u\right)^2}\sqrt{\sum_{i\in UPS_v}\left(\overleftarrow{w}_{v,i}-\overleftarrow{w}_v\right)^2}} \tag{7}$$

Among them, UPS_{uv} represents the set of preference types shared by users u and v, UPS_u and UPS_v represent the set of preference types of users u and v, respectively, and $w_{u,i}$ and $w_{v,i}$ represent the preference types of users u and v, respectively. The weights on i, $\overleftarrow{w}_u$, and $\overleftarrow{w}_v$ represent the average weights of all preference types of users u and v, respectively.

By calculating the similarity of any two users in the user space, the users meeting the present threshold φ are clustered, and the nearest neighbor communities with the same or similar preference types are obtained.

2.3. Smart recommendation

The main task of intelligent recommendation is to automatically recommend product items to target users based on the nearest neighbor community generated in Section 2.2. Set the target user m to be recommended, the user similarity sim, and search the nearest neighbors of the target user m in the entire user preference space UPS The user obtains the nearest neighbor set $\mathrm{UPS_{mi}} = \{\mathrm{UPS_1},\ \mathrm{UPS_2},\ \cdots, \mathrm{UPS_k}\}$ with the same or similar preferences as m, and for the preference type i, the similarity $\mathrm{sim(m,\ UPS_1)}$ between $\mathrm{UPS_1}$ and m is the highest, and the similarity between $\mathrm{UPS_2}$ and m is the highest. The similarity $\mathrm{sim(m,\ UPS_2)}$ is followed, and so on. The preference information of each user in $\mathrm{UPS_{mi}}$ is used for a weighted average to achieve the prediction of the target user's preferences and needs. The calculation formula is:

$$P_{m,i} = \overleftharpoon{w}_m + \frac{\sum_{u \in UPS_{mi}} sim(m,u) \times \left(w_{u,i} - \overleftharpoon{w}_u\right)}{\sum_{u \in UPS_{ui}} |sim(m,u)|} \tag{8}$$

Among them, $w_{u,i}$ represents the weight of user u for the preference type i, $\mathrm{sim}(m,u)$ represents the similarity between user m and u, $\overleftharpoon{w}_m$ and $\overleftharpoon{w}_u$ respectively represent users m and u in all. The average weight of the preferred type. The top n items with the highest predicted preference and demand are calculated by the above formula, that is, Top-N is recommended to the target user.

3. Algorithm experiment and analysis

To verify the effectiveness and efficiency of the collaborative filtering recommendation algorithm based on user preference mining designed in this paper, open-source data sets are used to carry out experimental design and performance verification of the algorithm.

3.1. Experimental data and evaluation criteria

The experimental data uses the Movielens data set ml-data collected by the GroupLens Research Group of the Department of Computer Science at the University of Minnesota in the United States. The data set involves a total of 17 types of movies, including action, adventure, war, animation, children, comedy, horror, science fiction, etc., including 943 users rated 100,000 records of 1682 movies, and user preference is mainly measured by the degree of rating. This article randomly selects 200 users' ratings of 400 movies from the data set and obtains the browsing of these users according to the log information of the website behavior.

The experimental evaluation index adopts the commonly used MAE (Mean Absolute Error, MAE) index and the Coverage index in the collaborative filtering algorithm to evaluate the efficiency and quality of the algorithm. The basic principle of the MAE index is to calculate the deviation between the user score and the actual user score. The accuracy of prediction, where the larger the MAE value, the worse the accuracy and quality of the recommendation; the smaller the MAE value, the better the accuracy and quality of the recommendation. The Coverage indicator is a widely used indicator for evaluating collaborative recommendation system recommendations The evaluation index of coverage refers to the coverage of user preferences of the item set recommended by the recommendation system for users. The greater the Coverage value, the stronger the recommendation system's ability to cover user preferences. The smaller the coverage value, the weaker the recommendation system's ability to cover user preferences.

3.2. Experimental design and result analysis

On the premise of the above experimental data and evaluation indicators, use the benchmark CF algorithm as a reference algorithm to conduct two sets of comparative experiments.

(1) The effect of the setting of the user similarity threshold φ on the execution result of the algorithm. By setting φ to a different value, set the number of nearest neighbors to K=20, calculate the benchmark CF algorithm and the ECFR-UPM designed in this paper The MAE and Coverage values of the algorithm, the experimental results are shown in Figure 2 and Figure 3.

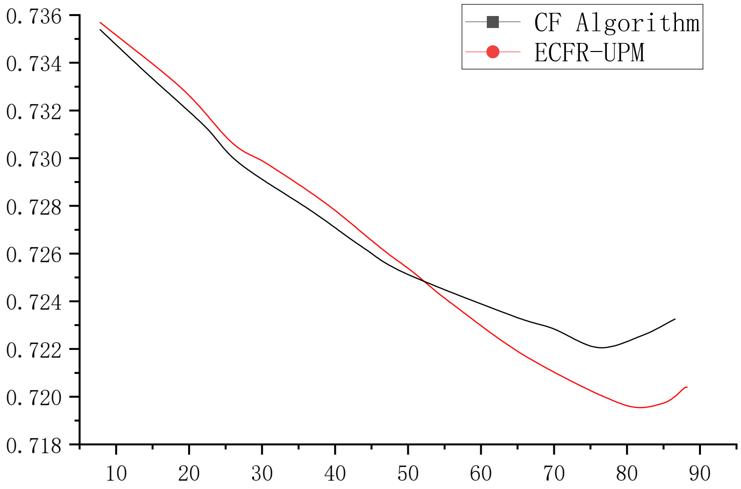

Figure 2. MAE comparison chart of different algorithms

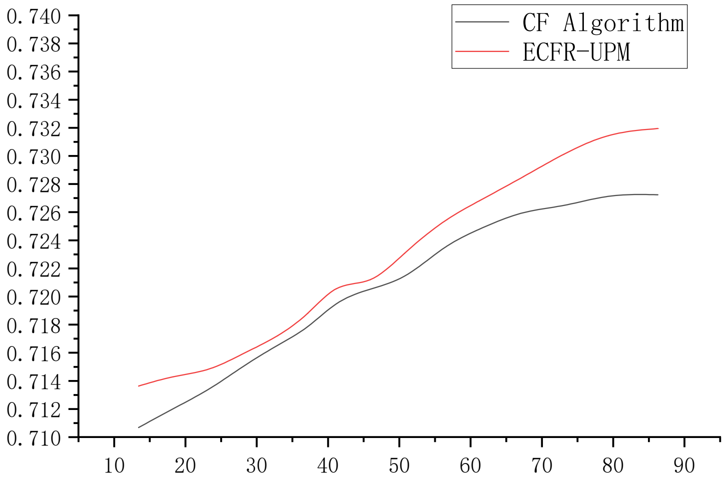

Figure 3. Coverage comparison chart of different algorithms

It can be seen from [Figure 2] and [Figure 3] that when the number of nearest neighbors is constant, the setting of the user similarity threshold φ has an impact on the MAE value and Coverage value of the two algorithms, and when φ = 80%, the two algorithms The MAE value of φ is at the lowest point, and the growth of the Coverage value tends to be balanced. However, regardless of the value of φ, the MAE of the ECFR-UPM algorithm is lower than that of the benchmark CF algorithm, especially when φ >40%, the gap is more obvious; At the same time, the growth of the Coverage value also shows that the ECFR-UPM algorithm is better than the benchmark CF algorithm.

(2) The effect of the setting of the number of nearest neighbors K on the execution results of the algorithm. By setting a different number of nearest neighbors K, set the user similarity threshold φ =80%, calculate the MAE and coverage of the benchmark CF algorithm and the ECFR-UPM algorithm designed in this paper Value, the experimental results are shown in the figure:

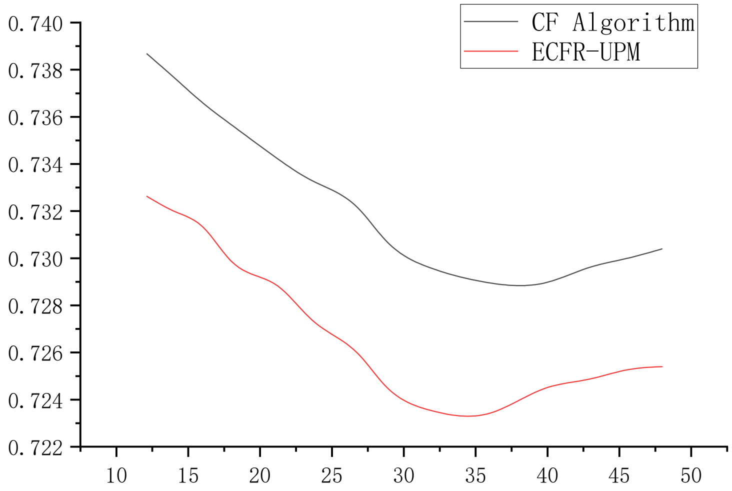

Figure 4. MAE comparison chart of different algorithms

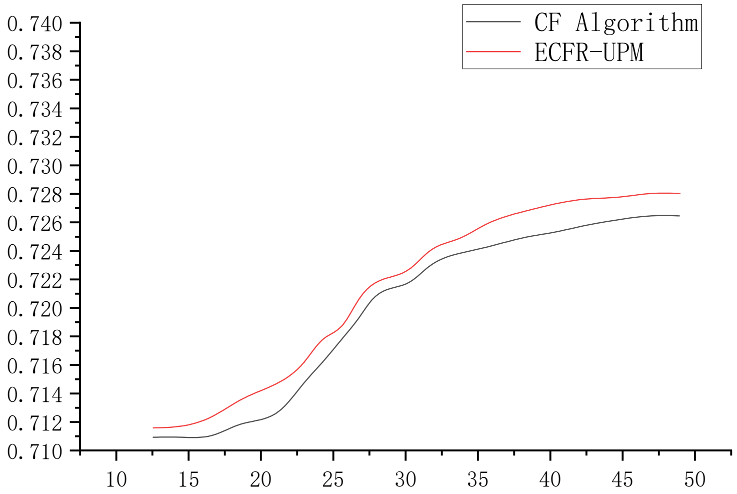

Figure 5. Coverage comparison chart of different algorithms

It can be seen from [Figure 4] and [Figure 5] that when the user similarity threshold φ is constant, the setting of the number of nearest neighbors has an impact on the MAE value and Coverage value of the two algorithms, and when K is between 35 and 45. The MAE value of the two algorithms is at the lowest point, and the growth of the coverage value tends to be balanced. In addition, although the setting of the K value affects the calculation results of the MAE value and the coverage value, on the whole, even at the best K Value range, the performance of the ECFR-UPM algorithm is significantly better than the benchmark CF algorithm.

Through the above two sets of comparative experiments, it can be seen that the ECFR-UPM algorithm designed in this paper has better performance than the benchmark CF algorithm on the two evaluation indicators of MAE and Coverage. This is mainly because the ECFR-UPM algorithm comprehensively utilizes user preferences. Knowledge is used to calculate user similarity and generate nearest neighbor communities. In the recommendation process, it combines user preferences and needs to make intelligent recommendations, which is conducive to improving the accuracy and quality of recommendations; while the benchmark CF algorithm only uses user registration and user ratings. Sexual user preferences are calculated for user similarity, and these explicit user preferences cannot fully and accurately reflect the real preferences and potential needs of users, so there are large errors in the actual application process.

4. Conclusion

Aiming at the problem that the traditional e-commerce collaborative filtering recommendation algorithm is difficult to accurately locate the nearest neighbor community of the target user and the recommendation quality and accuracy are not high, this paper designs an e-commerce collaborative filtering recommendation algorithm based on user preference mining. The algorithm uses users Preference mining technology to conduct the mining and analysis of users' explicit preference knowledge and tacit preference knowledge and realizes the construction of nearest neighbor community and intelligent recommendation based on user preference knowledge. Experiments show that the algorithm has achieved the expected results and comprehensively utilizes user preference knowledge for collaboration Filtering recommendation is the key to improving the accuracy and quality of recommendation results.

References

[1] Keunho, Choi, Donghee, “A hybrid online-product recommendation system: Combining implicit rating-based collaborative filtering and sequential pattern analysis” Electronic Commerce Research and Applications, **(2012)**

[2] H. Kim, J. K. Kim, and Y. Cho, “A collaborative filtering recommendation methodology for peer-to-peer systems,” International Conference on Electronic Commerce and Web Technologies. Springer, Berlin, Heidelberg, **(2005)**

[3] E. A. Kolchugina and V. A. Makar, “Method for scalable collaborative filtering recommendation systems,” **(2012)**

[4] B. M. Sarwar, G. Karypis, and J. A. Konstan, “Application of dimensionality reduction in recommender system: A case study,” Proceedings of the ACM WebKDDWorkshop, **(2000)**

[5] H. D. Ren and N. Jia, “Study on the classification for navigation query intention based on user similarity calculation, Journal of Xihua University (Natural Science Edition), **(2011)**

[6] P. Massa and B. Bhattacharjee, “Using trust in recommender system: An experimental analysis,” Proceedings of the 2nd Int’l Conf. on Trust Management, no.6, pp.221-235, **(2004)**

[7] H. Kim, J. K. Kim, and Y. Cho, “A collaborative filtering recommendation methodology for peer-to-peer systems,” International Conference on Electronic Commerce and Web Technologies. Springer, Berlin, Heidelberg, **(2005)**

[8] P. Kazienko and P. Kolodziejski, “Personalized integration of recommendation methods for e-commerce, International Journal of Computer Science and Applications, vol.3, no.3, pp.625-30, **(2006)**

[9] F. M. Madani and M. Memari, “A novel and optimized product recommendation method in e-commerce," Bilgi Ekonomisi Ve Ynetimi Dergisi, vol.2, no.1, **(2007)**

[10] H. R. Kim and P. K. Chan, “Learning implicit user interest hierarchy for context in personalization,” Applied Intelligence, vol.28, no.2, pp.153-166, **(2008)**

[11] B. Strunjas and Svetlana, “Algorithms and models for collaborative filtering from large information corpora,” The University of Cincinnati, **(2008)**

[12] Y. Wu and J. Zheng, “A collaborative filtering recommendation algorithm based on improved similarity measure method,” Proceedings of the 2010 IEEE International Conference on Progress in Informatics and Computing, **(2010)**

[13] T. Moriyasu and H. Tsuji, “Acquiring explicit knowledge for making evaluations from Tacit knowledge: Inductive inference of judgment rules from evaluation examples,” Journal of Japan Industrial Management Association, vol.62, no.3, pp.75-85, **(2011)**

[14] D. Godoy and A. Amandi, “Modeling interests of web users for recommendation: A user profiling approach and trends,” Evolution of the Web in Artificial Intelligence Environments, **(2008)**