Weighted interest degree collaborative recommendation algorithm based on association rules

Jiuzhi Lin[1]*, Zhaoxia Chang[2], and Jing Zhang[3]
[1]University of Science and Technology Beijing, Beijing 100083 China
lin_370681@163.com
[2]Minzu University of China, Beijing, 100081 China
S161187@muc.edu.cn
[3]Beijing Technology and Business University, Beijing, 100048 China
1172980129@qq.com

Abstract

With the development of the internet and mobile technology, the era of information overload has come, thereby leading to considerable interests in the recommendation system as an important mean to address such an overload problem. Especially, collaborative filtering recommendation algorithm is the most widely used and successful approach for the recommendation system. This paper first introduces the related concepts and principles of collaborative filtering, and then proposes a weighted interest model based on association rules to improve the accuracy of the proposed algorithm.

Keywords: association rules, interests, similarities

1 Introduction

In recent years, with the popularity of the Internet and the rapid development of information technology such as the Internet of Things and cloud computing, the amount of information contained in cyberspace has grown exponentially[9]. The vastness of network information resources has greatly improved the recall rate of information. However, the massive retrieval of information is accompanied by an increase in user search and screening time, and a decrease in the precision rate, which makes it difficult for users to find the information you want from a vast amount of network information resources in a short period of time. The explosive growth of network information has made the problem of “information overload” more serious[11]. A large amount of unrelated redundant data information seriously interferes with people’s choice of relevant useful information, making the cost of obtaining high-quality and valuable information higher and higher for users. The development of the recommended system has so far had a history of more than 20 years, due to its large application requirements, the recommendation system has received extensive attention[19]. As a filtering mechanism[15], the recommendation system is an important means to solve information overload[18]. The core of the recommendation system is the recommendation algorithm[17]. The traditional recommendation algorithm can be mainly divided into three categories: content-based recommendation algorithm [2], collaborative filtering recommendation algorithm [8, 14, 6], and hybrid recommendation algorithm [5, 13, 7].

The core idea of content-based recommendation algorithm is to extract attributes that can represent them from articles, news, commodities, etc., build project configuration files, and construct user configuration files through user’s behavior records and interests, Then compare the similarity between the

Research Briefs on Information & Communication Technology Evolution (ReBICTE), Vol. 4, Article No. 15 (November 15, 2018)

*Corresponding author: School of Computer & Communication Engineering, University of Science and Technology Beijing, Beijing, 100083, China project configuration file and the user profile, then recommend the item with the highest similarity to the user[10].

The collaborative filtering recommendation algorithm is currently the most widely used recommendation algorithm. The core idea is to analyze a user's interests and hobbies through the user's behavior record, then find users with similar interests to the target user from the user group, and then contact these users' items, news, songs, etc. that are not reached by the target user, but recommended to the target user.

The core idea of the hybrid recommendation algorithm is to combine multiple recommendation algorithms and process all the different algorithms so as to synthesize the results, or to combine multiple algorithms in different computing links so as to achieve faster and more accurate results of information push. This paper first introduces the related concepts and principles of collaborative filtering, then proposes a weighted interest model based on association rules, improves the accuracy of the proposed algorithm, and finally summarizes the article.

This paper first introduces the related concepts and principles of collaborative filtering, and proposes a weighted interest model based on association rules. Then, the accuracy of the proposed algorithm is improved, followed by the summary.

2 Collaborative filtering recommendation

The core idea of collaborative filtering algorithm is to find neighbors based on similarity, and then according to the prediction scores and recommendations, the process can be divided into the following three parts: collect data, calculate similarity, rank the ratings, and make recommendations[4, 3].

2.1 Similarity calculation

The purpose of the collaborative filtering algorithm is to use the similarity calculation to score the unrated items based on the user u's scoring items. The computation of similarity between users becomes one of the keys to collaborative filtering algorithms. Commonly used similarity measures include Euclidean distance, Pearson correlation coefficient[1], cosine similarity [16], modified cosine similarity [12], and so on.

The similarity is calculated in two ways. One is based on the similarity of items, and the other is based on the user's similarity. The choice between the two usually depends on the number of users or items. The following uses the user-based collaborative filtering as an example to introduce the following similarity calculation methods.

2.2 Euclidean distance

Constructing a rating matrix based on user ratings, setting the row vectors to represent different users' ratings, and the column vectors to different users' ratings of the same item. The similarity is represented by Formula (1):

$$sim = \frac{1}{1+d_{AB}} \tag{1}$$

Among them, *sim* indicates the degree of similarity, which ranges from 0 to 1, and d_{AB} represents the Euclidean distance of the score of the two items, and $d_{AB} = \sqrt{(a_1-b_1)^2+(a_2-b_2)^2+\cdots+(a_n-b_n)^2}$.

2.3 Pearson correlation coefficient

The Pearson correlation coefficient can be used to measure the similarity between two vectors. Obviously the advantage of this method over Euclidean distance is that it is insensitive to user ratings, such as user *a* scoring 5 points for all items. While user *b* scores 1 for all items, the Pearson correlation coefficient considers the two vectors to be equal. Equation (2) shows the Pearson correlation coefficient similarity:

$$sim\left(a,b\right)=\frac{\sum\limits_{i\in I_{ab}}\left(r_{a,i}-\overline{r_a}\right)\left(r_{b,i}-\overline{r_b}\right)}{\sqrt{\sum\limits_{i\in I_{ab}}\left(r_{a,i}-\overline{r_a}\right)^2}\sqrt{\sum\limits_{i\in I_{ab}}\left(r_{b,i}-\overline{r_b}\right)^2}} \tag{2}$$

Among them, I_{ab} denotes a set of items that all users *a* and *b* have jointly scored, $r_{a,i}$, $r_{b,i}$ respectively denotes the ratings of user *a* and *b* for item *i*, $\overline{r_a}$, $\overline{r_b}$ respectively denotes the average rating of user *a* and *b*.

2.4 Cosine similarity

Build a score matrix based on user ratings, set the row vector to represent different users' ratings, and the column vector indicate different users' ratings for the same item. The degree of similarity is measured by calculating the cosine of the angle between two vectors. If the two vectors have the same direction, the similarity is 1.0; if the angle is 90 degrees, the similarity is 0. Its formula is shown in equation (3):

$$\cos\theta=\frac{A\cdot B}{\|A\|\,\|B\|} \tag{3}$$

Among them, *AB* represent two scoring vectors of two items, $\|A\|$, $\|B\|$ represent the 2 norm of the vector.

2.5 Modified cosine similarity

Since the cosine similarity measurement method does not consider the rating criteria of different users, people proposed a modified cosine similarity calculation method. The calculation formula is shown in equation (4):

$$sim\left(a,b\right)=\frac{\sum\limits_{i\in I_{ab}}\left(r_{a,i}-\overline{r_a}\right)\left(r_{b,i}-\overline{r_b}\right)}{\sqrt{\sum\limits_{i\in I_{a}}\left(r_{a,i}-\overline{r_a}\right)^2}\sqrt{\sum\limits_{i\in I_{b}}\left(r_{b,i}-\overline{r_b}\right)^2}} \tag{4}$$

Among them, I_{ab} denotes a set of items that all users *a* and *b* have jointly scored, I_a and I_b respectively denotes the collection of individual scoring items,$r_{a,i}$ and $r_{b,i}$ respectively denotes the rating of the user *a* and *b* for the project *i* , $\overline{r_a}$, $\overline{r_b}$ respectively denotes the average score of user *a* and *b*.

3 Association-based collaborative recommendation algorithm based on weighted interest model

3.1 Construction of user interest model

The user interest model is a model representation of user information requirements and is the core of a personalized service recommendation system. The user model can usually show the user's interest in some specific topic information, which provides a basis for service providers to provide customers with more convenient services. The user model can usually be established in two ways. One is to directly obtain the user's interest and information demand tendency through the method provided by the user; the other method is to track and analyze the user's search, viewing, and other behavior records through the system to build a user profile.

In order to provide users with better services, this paper adopts a combination of active and passive acquisition to establish a user interest model. First of all, when actively acquiring user interests, the user can select the tags he is interested in from the list and perform the rating (the rating ranges from 1 to 5 stars) to express his/her own preference for the tag items, that is, the weight of interest. When the label in the table cannot meet the user's needs, the user can add keywords that he is interested in by manually inputting the information, and the user can also change the interest label and the score at any time and any place. The user scoring matrix is constructed by this method of actively acquiring interest tags and interest scores.

3.2 Association rule mining algorithm

Association rules can be used to find meaningful internal relationships in large-scale data sets. These relationships can take two forms: frequent item-sets and association rules. Frequent item-sets are collections of items that often appear together, and association rules suggest that there may be strong associations between the two items. The most famous case in the association analysis is the "beer diapers case" in the supermarket "shopping basket" data. Through the analysis of "shopping basket" data, customers' consumption habits can be known. Assuming that $I = \{i_1, i_2, ..., i_m\}$ is a set of m different data items, where the element is called items, and the set of items is called item-sets. Assuming that $D = \{T_1, T_2, ..., T_n\}$ is a transaction database, and each transaction is a subset of item-sets, then$|D|$ represents total number of transactions D. Implicative formula of association rule is shown in Formula (5).

$$R : X \Rightarrow Y \tag{5}$$

Among them,$X \subset I$, $Y \subset I$,and $X \cap Y = \emptyset$, item-set X appears in a transaction, which leads to Y also appears in a transaction with a certain probability. Analysis of association rules for user interest tags can be measured by two criteria: support and credibility.

Support indicates the probability that the item-set {X,Y} appears in the total item set, that is, the ratio of the number of items in the item set to the total number of item-sets. The formula is shown in Formula (6).

$$Support(X \Rightarrow Y) = P(XY) = \frac{\text{sum}(\text{X}\bigcup\text{Y})}{|D|} \tag{6}$$

Confidence indicates the probability of the item-set containing Y in the item-set containing X. The formula is shown in Formula (7).

$$Confidence(X \Rightarrow Y) = P(X|Y) = \text{sum}(\text{X}\bigcup\text{Y})/\text{sum}(\text{X}) \tag{7}$$

3.3 Frequent Itemset Mining Algorithm

In the specific application of association rule mining, the efficiency of the algorithm is undoubtedly very important, which is also the research focus of current data mining. In the process of data mining, the frequent item-set generation process involves a huge amount of calculation, which is the most difficult and complexity part. Currently, there are three frequent item-sets mining algorithms commonly used: Apriori algorithm, Eclat algorithm, and FP-Growth algorithm. The following is a brief description of the Apriori algorithm as an example.

Apriori principle: If a set of items is frequent, then all its subsets are also frequent. From this we can infer that if a set of items is an infrequent item-set, then all its supersets are also infrequent, and through this principle we can reduce the computation time and increase the efficiency of the algorithm. In the process of execution, the Apriori algorithm first finds all the frequent 1-item sets by scanning all the databases, then finds all the candidate 2-item sets through the Apriori_Gen algorithm, and then counts each item set to find all the frequent 2 – Item-sets... By analogy, frequent k-item sets can always be found in an iterative manner.

3.4 Weighted interest model based on association rules

In the first part of this section, we obtain interest information by user-actively setting tags and scoring, and establish a user interest scoring matrix. However, in the process of actively setting the label, due to the inaccurate expression of the interest tag by the user, the simple and straightforward way cannot construct an accurate user model; therefore, based on this, we introduce an association rule algorithm to deeply study the relationship of user interest. Here we assume $I = \{i_1, i_2, ..., i_m\}$is a collection of interest tags made up of m different tag items. And we assume that $D = \{T_1, T_2, ..., T_n\}$ is a character database, where each character is a subset of item-sets. Here, we add various tags to each transaction, such as: sports, we can set the following tags (basketball, football, tennis, table tennis, etc.), through which we can better get more accurate the interest of users, so as to build user interest model, $\{(t_1, w_1), (t_2, w_2), \cdots (t_n, w_n)\}$, $t_1, t_2, \cdots, t_n$ represent different interest tags, and $w_1, w_2, \cdots, w_n$ represents the weight of the corresponding interest label. According to the user rating matrix, we find the n tag items with the highest similarity to the target user a, and generate the nearest neighbor set of users $B_a = \{b_1, b_2, \cdots, b_n\}$, According to Formula (8), we can predict the user's prediction score for the interest label item.

$$p_{a,i} = \overline{R_a} + \frac{\sum\limits_{b \in B_a} \left(R_{b,i} - \overline{R_b}\right) sim(a,b)}{\sum\limits_{b \in N_a} |sim(a,b)|} \tag{8}$$

In the equation, $sim(a,b)$ represents the similarity between a and b, and $R_{b,i}$ represents the score of the users' interest tag i in the nearest neighbor set $B_a = \{b_1, b_2, \cdots, b_n\}$,$\overline{R_a}$ and $\overline{R_b}$ represent the average scores of users $a$and b their respective interests scores. Based on the predicted value of interest labels and the users' interest model $\{(t_1, w_1), (t_2, w_2), \cdots (t_n, w_n)\}$obtained from association analysis, We can get a more accurate prediction score through Formula (9):

$$f_{a,i} = w_i \cdot p_{a,i} = w_i \left(\overline{R_a} + \frac{\sum\limits_{b \in B_a} \left(R_{b,i} - \overline{R_b}\right) sim(a,b)}{\sum\limits_{b \in N_a} |sim(a,b)|} \right) \tag{9}$$

3.5 Experimental data

Since this article uses the combination of active and passive methods to obtain the user's interest score, but most of the current rating information in the Internet is the score after shopping and watching. It is not applicable to the experimental test of this article. Therefore, this article obtains some students' interest ratings for various news through questionnaire survey. The content of the questionnaire includes military, finance,entertainment, technology, digital, history, sports, movies, etc, with a score range of 1-5, and detailed classification of each content. Such as sports are divided into: basketball, tennis, football, table tennis and so on.

3.6 Simulation experiment and result analysis

From Figure 1 we can see that boys who like to watch action movies usually prefer to focus on basketball, and girls who like action movies usually prefer badminton. This experiment uses Mean Absolute Error (MAE) to measure the accuracy of the proposed algorithm. MAE is a commonly used measurement method for measuring the accuracy and comparison of statistics and can accurately reflect the quality

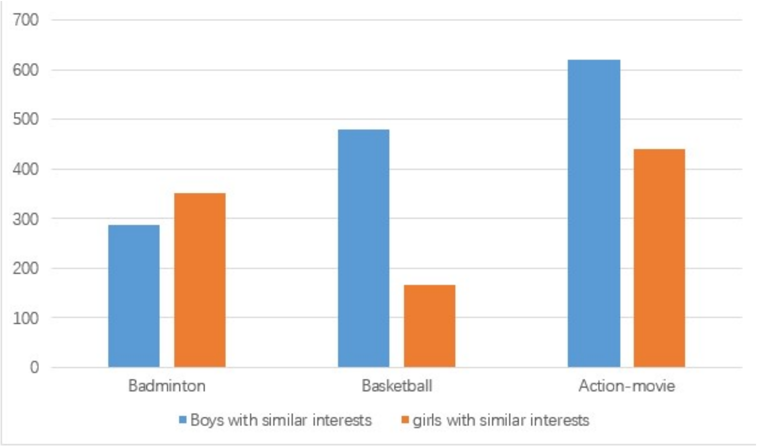

Figure 1: Some Interest Association Diagrams

of recommendation. [14] It can be used to measure the predicted user rating and the actual user rating. The smaller the error, the smaller the value of the MAE, the higher the accuracy of the recommendation. Conversely, the accuracy of the recommended algorithm is worse. The calculation method is shown in formula 2.6. (10).

$$MAE = \frac{\sum_{i=1}^{n} |R_{a,i} - f_{a,i}|}{n} \tag{10}$$

In the formula, $R_{a,i}$ represents the actual score of interest tag i by user$a$, and $f_{a,i}$ is the predicted score of interest tag i by user$a$ predicted by this algorithm. The MAE value comparison between the proposed algorithm and the traditional recommendation algorithm (using only the similarity calculation formula) is shown in Figure 2. In the figure, the vertical axis represents MAE value, horizontal axis indicates the number of users of the neighboring matrix used when predicting the score.

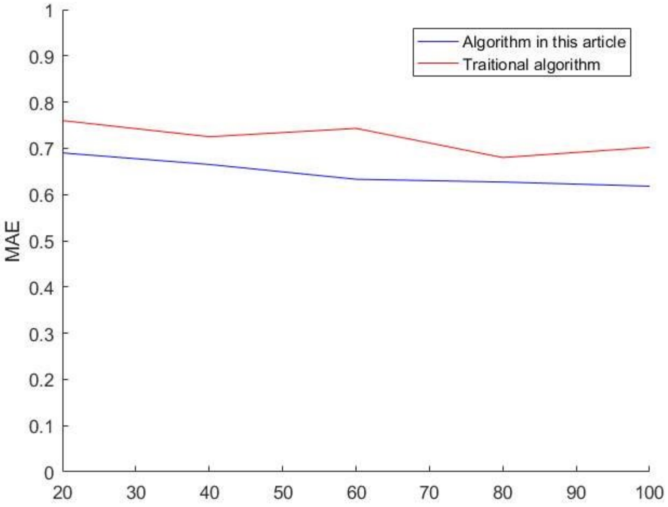

Figure 2: MAE value comparison

From Figure 2, we can see that with the increase of the number of neighbor matrix users, the MAE value of the proposed algorithm is constantly decreasing, and the overall MAE value of the proposed algorithm is lower than that of the traditional recommendation algorithm. This shows that the proposed algorithm is superior to the traditional recommendation algorithm.

4 Summary and outlook

This paper proposes a weighted interest user model based on association rules. It uses the association rule to mine the user's interest label to calculate its weight value, and uses the scoring matrix to accurately calculate the user's nearest neighbor set, thereby obtaining a more accurate interest score and improving the accuracy of the recommendation. However, in the interest model of this paper, data sparsity [20] and scalability in collaborative recommendation systems are not fully considered. In the future work, we will further study the solution to the challenges presented by the data sparsity and scalability issues in the recommendation system.

5 Acknowledgements

We would like to express our sincere thanks to the referees and the editors for their valuable suggestions. This work is supported by the Foundation of Science and Technology on Information Assurance Laboratory No.KJ-17-101.

References

[1] M. Alshamri and N. Alashwal. Fuzzy-weighted similarity measures for memory-based collaborative recommender systems. *Journal of Intelligent Learning Systems and Applications*, 6(1):42566:1–42566:10, Februay 2014.

[2] G. Arora, A. Kumar, G. Devre, and A. Ghumare. Movie recommendation system based on users' similarity. *International Journal of Computer Science and Mobile Computing*, 3(4):765–770, April 2014.

[3] Y. Bergner, S. Droschler, and K. G. Model-based collaborative filtering analysis of student response data: Machine-learning item response theory. In *Proc. of the 5th International Conference on Educational Data Mining (EDM'12), Chania, Crete, Greece*, January 2012.

[4] M. Brooks and H. Carsten. listings: Typeset source code listings using LaTeX. `http://www.ctan.org/tex-archive/help/Catalogue/entries/listings.html` [Online; accessed on May, 2008], 1986-2006.

[5] R. Burke. Hybrid recommender systems: Survey and experiments. *User Modeling and User-Adapted Interaction*, 12(4):331–370, November 2002.

[6] Y. Cai and H. Leung. Typicality-based collaborative filtering recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 26(3):766–779, March 2014.

[7] W. Chen and Z. Niu. A hybrid recommendation algorithm adapted in e-learning environments. *World Wide Web*, 17(2):271–284, March 2014.

[8] M. Ekstrand, J. Riedl, and J. Konstan. Collaborative filtering recommender systems. *Foundations and Trends in Human-Computer Interaction*, 4(2):81–173, February 2011.

[9] G. George, M. Haas, and A. Pentland. Big data and management. *Academy of Management Journal*, 57(2):321–326, April 2014.

[10] S. Gong and H. Ye. Combining memory-based and model-based collaborative filtering in recommender system. In *Proc. of the 2009 Pacific-Asia Conference on Circuits, Communications and Systems (PACCS'09), Chengdu, China*, pages 690–693. IEEE, May 2009.

[11] R. H. Gouws and S. Tarp. Information overload and data overload in lexicography. *International Journal of Lexicography*, 30(4):389–415, July 2016.

[12] X. He and Y. Luo. Mutual information based similarity measure for collaborative filtering. In *Proc. of the 2010 IEEE International Conference on Progress in Informatics and Computing (PIC'10), Shanghai, China*, pages 1117–1121. IEEE, December 2011.

[13] V. Kagita and A. Pujari. Virtual user approach for group recommender systems using precedence relations. *Information Sciences*, 294:15–30, February 2015.

[14] G. Linden, B. Smith, and J. York. Amazon.com recommendations. *IEEE Internet Computing*, 7(1):76–80, January 2003.

[15] V. A. Patil and L. Ragha. Comparing performance of collaborative filtering algorithms. In *Proc. of the 2012 International Conference on Communication, Information & Computing Technology (ICCICT'12), Mumbai, India*, pages 1–6. IEEE, October 2012.

[16] G. Pirlo and D. Impedovo. Cosine similarity for analysis and verification of static signatures. *IET Biometrics*, 2(4):151–158, December 2013.

[17] A. Popescul, D. Pennock, and S. Lawrence. Probabilistic models for unified collaborative and content-based recommendation in sparse-data environments. In *Proc. of the 7th Conference on Uncertainty in Artificial Intelligence (UAI'01), Seattle, Washington, USA*, pages 437–444. Morgan Kaufmann Publisher Inc., August 2001.

[18] F. Ricci, L. Rokach, and B. Shapira. *Recommender Systems Handbook*. Springer US, 2015.

[19] B. Sarwar and G. Karypis. Item-based collaborative filtering recommendation algorithms. In *Proc. of the 10th International Conference on World Wide Web (WWW'01), Hong Kong, Hong Kong*, pages 285–295. ACM, May 2001.

[20] X. Yu. and L. Min-Qiang. Effective hybrid collaborative filtering algorithm for alleviating data sparsity. *Journal of Computer Applications*, 29(6):1590–1593, July 2009.

Author Biography

Jiuzhi Lin is a master student in the University of Science and Technology Beijing, majoring in Electronics and Communication Engineering. His research interest is edge computing.

Zhaoxia Chang is a 2019 graduate of the upcoming master's degree, and now she is studying at the Minzu University of China, majoring in Computer Science and Technology.

Jing Zhang received her master's degree in quantitative economics from Beijing technology and business university in June 2018. Since January 2018, she has been working for Jingdong century trading co., LTD with focus on the area of data analysis and mining.