A Collaborative Filtering Algorithm based on Citation Information

Tian Bai
College of computer Science and Technology
Jilin University
Changchun, China
baitian@jlu.edu.cn

Binzhao Ding
College of computer Science and Technology
Jilin University
Changchun, China
609718685@qq.com

Ye Wang
College of computer Science and Technology
Jilin University
Changchun, China
yewang12@mails.jlu.edu.cn

Jingbo Ning
College of computer Science and Technology
Jilin University
Changchun, China
975618327@qq.com

Lan Huang *
College of computer Science and Technology
Jilin University
Changchun, China
huanglan@jlu.edu.cn
* Corresponding Author

Abstract—Objective: With the rapid growing number of published scientific papers in the age of big data, users often find themselves difficult to select useful information from such massive academic information. This paper aims at the problems of collaborative filtering techniques in scientific citation data.

Methods: This paper proposes an improved machine learning algorithm, that is designed to predict user ratings of academic theses by using Fisher Linear Regression combined with information of confidence scores, preference scores, the number of cooperative users in science citation data and the actual citation scores.

Results: Multiple features considered in the algorithm have a positive impact on the recommendation results.

Conclusion: This paper proposed a collaborative filtering recommendation algorithm combined with citation data analysis in order to improve the accuracy of predicted results. The experiments have shown that the proposed algorithm is proved to be effective and improve the accuracy of recommendation.

Keywords- *Collaborative Filtering; Citation Information; Citation Network; Recommendation System; Fuzzy Clustering*

I. INTRODUCTION

When facing the massive academic information, scientific researchers need an access to locate their interested papers accurately and conveniently, such as a platform of personal information recommendation. Therefore, how to realize such platform to obtain users' interested information by using data mining and machine learning methods is discussed in this paper.[1,2,3,4]

At present, analysis in this area is largely focused on papers written in English. International famous citation analysis tools such as DBLP[5], Microsoft Academic Search[6], Google Scholar are most frequently used. In particular, papers are sorted by the citation number combined with level of journal or conference to ensure that the most valuable papers having the highest precedence to show in Google Scholar. Top-N of Google Scholar is the early form for recommendation. DBLP tends to be engaged in author search in the fields of computer science.

Based on Fuzzy Similar RM to balance the accuracy and novelty of papers, this paper obtains user confidence score. At the same time, based on the method of machine learning, this paper proposes "paper-feature" matrix to solve the problem of user preference score. Based on Linear Integration, papers are recommended by weighting the confidence score and preference score linearly.

II. RELATED WORK

Content-based recommendation [7] is to recommend new items with similar characteristics with the items already rated by the user. Collaborative filtering is proposed by Goldberg et al. The description of the algorithm is: first, calculate the preference of similar users; then, process and filter all kinds of information on the above basis. Different recommendation methods have different advantages. Thus, it is better to mix different technology in recommendation to take their advantages and overcome their shortcomings.

The essence of the matrix decomposition based on collaborative filtering recommendation algorithm is to filter the score matrix of User-Item. There are many ways to filter the matrix. In order to make sure the selected methods have the minimal impact on the original matrix, the characteristic values of matrix must be considered. It is generally accepted that the filtering method has little influence on the original matrix when the characteristic values of the complementary matrix are close to those of original matrix, for example, the method of Singular Value Decomposition.

Traditional method of science citation nodes similarity based on formula is simple and has low time complexity, but the disadvantage is also obvious. In 2010, Ji Liu proposed a cooperative network community analysis recommendation method [8] from the perspective of community detecting, but the emergent novelty was not adequately considered. In 2008, Yong Chen proposed an improved algorithm based on similarity [9], but did not take into account characteristics of big data sets, thus the performance and processing efficiency are considerably affected.

Fuzzy clustering is an algorithm that has advantages of novelty, precision, and efficiency of recommendation. The algorithm can obtain different matrices from data of difference topology structure, and use these matrices on the corresponding data for further analysis and predication. In addition, the results of fuzzy clustering can be more accurate if the values calculated based on formula are taken as a feature. The next sections propose solutions to issues of data set size, novelty and precision of recommendation.

III. RMBCI BASED COLLABORATIVE FILTERING ALGORITHM

A. *Fuzzy Clustering based Confidence Score Algorithm*

1) *The Basic Idea*

First, in the concept of RMBCI (Rate Matrix based on Citation Information), a paper citer is taken as a User; a cited paper is taken as an Item; the number of citation is taken as the Rate by users to the item. Second, the fuzzy concept is added to the RMBCI, based on which, a Fuzzy Similar Rate Matrix named RMBFS (RMBFS, Rate Matrix Based on Fuzzy Similar) can be proposed. User groups are established using fuzzy clustering of IKHAFC (Improved KHA Fuzzy Clustering). Finally, we can calculate the confidence score according to the clustering of IKHAFC.

2) *Fuzzy Clustering Algorithm*

FMBFS based Fuzzy Clustering Steps:

Ite: The maximum iteration number; k: cluster number; e: threshold; E: the disparity between two of nearest data

(a) Initialize the fuzzy matrix U by generating n*k random value between 0 and 1;

(b) Calculate the distance between attributes for each category;

(c) Calculate all cluster centers Q;

(d) Calculate the fuzzy matrix U;

(e) If E is less than e or Ite is equal to 0, the algorithm is over; else Ite = Ite-1, go to (b).

3) *Generate Confidence Score*

First, on the basis of RMBFS, results of science citation data are calculated; then, the confidence score is calculated using the User-based Collaborative Filtering.

Let v be the neighbor of user u, $\mathrm{sim}(u, v)$ represents the similarity between user u and user v, is the score of item i rated by user u. Thus, the score of item i rated by user u can be calculated as follows:

$$r_{ui} = \sum_{v \in \Gamma(u)} s\mathrm{im}(\mathrm{u,v}) r_{vi}$$

Confidence score from the science citation data is obtained only using the information of citations by users. Experiment results shows that the method can achieve good results. It is necessary to combine the score of this method with the score of collaborative filtering recommendation algorithm to form the final score.

B. *RMBCI based Preference Score Algorithm in Collaborative Filtering Recommendation*

1) *Main Idea of the Algorithm*

Because of time and space limits, it is difficult for some of traditional collaborative filtering algorithms to perform in the case of big data. For example, there are millions of papers in Microsoft Academic. If the user-based collaborative filtering recommendation algorithm is directly used, time - complexity is O (N2M) in the calculation step of user similarity, where N is user, M is item. The time -complexity is higher than accepted level. The space complexity can also be very large using the collaborative filtering algorithms such as time or user-based model. Besides, both of user-based and item-based collaborative filtering algorithms have limit capacity on recommendation to new users.

Based on above discussion, it is necessary to propose an improved collaborative filtering algorithm to solve problems above. Aiming at the characteristics of big data, this paper proposes SVD algorithm based on user feature-paper (SVD_RBUFP). Preference Score is calculated according to such algorithm. Finally, the linear fusion method offers a subtle blend of confidence score charm with preference score to form the final score.

The description of the algorithm is: the structure is a User-Paper matrix, $R \in R^{n \times m}$ where n is the number of users, m is the number of papers. r_{fi} is the score that is rated by the users of feature f in the matrix. If the number of users possessing feature f is less than a given threshold, r_{fi} is zero to make it credible. And this kind of r_{fi} is what the improved collaborative filtering recommendation algorithm needs to solve.

SVD model is selected because SVD can be effective in reducing dimensions of data, and it integrates characteristics of paper into the model. Thus, SVD may be a more desirable alternative.

The formula of predicting model:

$$\hat{r}_{fi} = \mu + b_f + b_i + \sum_{c} p_{uc} q_{ic}$$

where μ is general score, b_f is the bias of user's feature, b_i is bias of paper feature and c is latent semantic.

$p_{uc} = P(u,c)$, $q_{ic} = Q(i,c)$, P and Q are the matrices after decomposition. SVD minimizes the RMSE to learn the matrices of P and Q using the training set.

Suppose a User has a set of Features, defined as $F(u)$. After calculating the Feature-Paper score matrix with the method of SVD, we then calculate the score of User-Paper by each user. In this paper, the formula of the linear weighted fusion method is as follows:

$$r_{ui} = \sum_{f \in F(u)} w_f r_{fi}$$

2) The Selection of User Feature

The selection of user feature is an important step of the improved collaborative filtering algorithm. Generally speaking, if the number of user features is below certain level, the model can't distinguish and describe users. Also, the number of user features should not be too high, or it will result in matrix sparse and high complexity.

The next section illustrates some key features of users. The Microsoft Academic Data Set is used as an example.

Feature 1: Age Feature

The feature of age (AF, Age Feature): user age is a natural number. Users are categorized according to age in 10 years. There are seven groups with the number of group to be age feature of a user. Group 1: in the age-bracket 20-30; Group 2: in the age-bracket 31-40; Group 3: in the age-bracket 41-50; Group 4: in the age-bracket 51-60; Group 5: above 60; Group 6: abnormal group; Group 7: the group without age.

Feature 2: Research Depth Feature

The feature of research depth (RDF, Research Depth Feature): there are four groups: Group 1: the bachelor group; Group 2: the master group; Group 3: the Doctor group; Group 4: the Postdoctoral group.

Feature 3: Activity Feature

The feature of user activity (AF, Activity Feature): the number of papers published by a user represents his activity in academic. There are five groups. Group 1: in the paper-bracket 1-3; Group 2: in the paper-bracket 4-10; Group 3: in the paper-bracket 11-30; Group 4: above 30; Group 5: no paper.

Feature 4: Keyword Feature

The feature of keyword (KF, Keyword Feature): keywords are included in citation data sets in order to facilitate users. The keywords are from the title of a paper. It is possible to extract the research direction, algorithm, innovation points, etc. Implied Dirichlet Allocation (IDA) method is used to analyze classification of potential user. According to interest vector of potential user, the classification with the highest weight is selected.

Feature 5: Feature Combination

FC (Feature Combination) is the common method to generate features in Machine Learning (ML). Although this method increases the dimension, it can deal with most of the related features effectively. There are ten combination features in the paper: AC×RDF，AC×AF，AC×KF，RDF×AF，RDF×KF，AF×KF，AC×RDF×AF，AC×RDF×KF，AC×AF×KF，RDF×AF×KF.

3) Calculate preference score

The specific steps of SVD-RBUFP method are given:

(1) Calculate all features of each user.

(2) Construct the matrix with Feature-Paper of user: compute the score of feature- paper rated by user; filter the item of which score is less than a specified threshold.

(3) Solve the matrix of feature-paper with the method of SVD.

(4) Calculate user preference score with the formula of linear weighted fusion according to the user features and feature-paper vector.

C. Confidence Score and Preference Score

In this paper, the results of Confidence Score and Preference Score are inputs of fusion algorithm analyzed by logistic regression. That is, we integrate confidence score charm with preference score to form the final score to make recommendation. The most frequently used method is Linear Integration method:

$$r_{ui} = \partial r_{ui}{}^{trust} + (1-\partial) r_{ui}{}^{preference}$$

where, $r_{ui}{}^{trust}$ is Confidence Score of user u to paper i, $r_{ui}{}^{preference}$ is Preference Score of user u to paper i and ∂ is the parameter of weight. We can adjust the weight between Confidence Score and Preference Score by ∂.

Though Linear Integration (LI) method is simple, it has obvious disadvantages: it may neglect some factors that affect the final score, for example, the number of cooperators and the number of features. The more the cooperators are, the larger the weight of Confidence Score is. Also, the more the features are, the more credible the weight of Preference Score is.

In this paper, Fisher Linear Regression (FLR) method is selected to integrate Confidence Score and Preference Score in order to avoid the disadvantages and limitations of Linear Integration in the actual environment. The features of FLR include: User Trust Score(UTS), User Preference Score(UPS), User Cooperator Number(UCN), User Feature Number(UFN) and so on. The formula of FLR is given as below:

$$r_{ui} = \partial r_{ui}{}^{trust} + (1-\partial) r_{ui}{}^{preference} + \lambda k_u + \theta f_u + \varphi k_u r_{ui}{}^{trust} + \phi f_u r_{ui}{}^{preference}$$

The approach of FLR allows the parameters of ∂ , β , λ and θ to be well defined. Experiments have shown that FLR is better than LI in the fusion of Confidence Score and Preference Score.

IV. EXPERIMENTS

A. data sets

DLB and Microsoft Academic provide citation data sets that can be used in proposed collaborative filtering blending recommendation algorithm.

B. evaluation criteria

RMSE (Root Mean Square Error) and MAP (Mean Average Precision) are adopted in the experiment.

RMSE is defined as follows:

$$RMSE = \sqrt{\frac{\sum_{u,i \in T} (r_{ui} - \hat{r}_{ui})^2}{|T|}}$$

where, r_{vi} is the actual score of item i rated by user u and $\hat{r}_{ui}$ is the predictable score.

Supposing that there are m sorted recommendation lists, a user may click one or more, or do nothing. When n recommendations are selected, the mean average precision is:

$$map@n = \sum_{k=1}^{n} P(K) / c$$

where c is the hits of the papers by a user. If c is 0, the result of the map is 0. $P(K)$ is the precision degree until the Kth paper in the list ($P(K)$ =number of cited paper until the Kth paper / K). If the Kth paper is not in the recommendation list, $P(K)$ is 0. The default value of n is 3 in the recommendation system. Mean Average Precision of N users is:

$$MAP@n = \sum_{i=1}^{N} map@n_i \; / \; N$$

C. *Experimental Results and Analysis*

Experiment A:

After the matrix of RMBFS is obtained, 8541 positive samples and 8541 negative samples are taken out from the papers of which citation is less than 169. Based on this, similarity based formula calculation method and fuzzy clustering algorithm method is used to make predictions. The predict accuracy of different similarity based methods is shown in Table 1. The table shows that the predict accuracy of FMBFS is obviously higher than other traditional methods.

TABLE I. PREDICTION ACCURACY OF DIFFERENT SIMILARITY BASED METHODS

Algorithm	accuracy
Common Neighbor(CN)	62.81%
Jaccard Index(JI)	64.34%
Salton Index(SI)	65.55%
Priority Link Index(PLI)	71.87%
Adamic-Adar Index(AAI)	72.31%
Resource Allocation Method(RAM)	72.89%
Logistic Regression(LR)	74.22%
Rate Matrix Based on Fuzzy Similar(FMBFS)	77.12%

Experiment B:

2000 data pairs from DBLP are selected in this experiment to form the RMBUFP (Rate Matrix on User Feature-Paper). Different methods for solving matrix will result in different RMSE, as shown in Table 2.

TABLE II. DIFFERENT RMSE OF DIFFERENT METHODS FOR SOLVING MATRIX

Solving Method	RMSE
Global Mean Model(GMM)	0.06533
Mu + bi + bf(MIF)	0.04954
Latent Semantic Model(LSM)	0.04247
SVD Recommend based on User Feature-Paper(SVDRUFP)	0.03923

In Table 2, the first method is using the global average value (M_u) to complement the matrix. The second method is using the global average value (M_u), bias of user feature B_f and bias of paper B_i to complement the matrix. The third method is Latent Semantic Model. The last method is SVDRUFP with better performance than other methods. So, creating feature-paper matrix followed by the solution of matrix decomposition model is better suited for recommendation.

Experiment C:

Different recommendation methods are used to accurately predict paper score rated by users. Then, the algorithm is evaluated on the local machine. Finally, the results of these algorithms and reliability analysis are given.

TABLE III. EVALUATION RESULTS OF DIFFERENT RECOMMENDATION ALGORITHMS

Recommendation Algorithm	RMSE	MAP@3
Naive Bayes	0.82629	0.41253
M-SVD	0.73686	0.39091
M-SVD+SNA CF(1)	0.65435	0.37379
M-SVD+SNA CF(2)	0.65064	0.39031

Naive Bayes uses simple Bayesian as the benchmark. M-SVD is an improved collaborative filtering recommendation algorithm. M-SVD+LI-CF (1) is the simple Linear Integration (LI) method which is used to integrate Confidence Score and Preference Score. M-SVD+FIR-CF (2) method is the Fisher Linear Regression (FLR) which is used to blend of confidence score charm with preference score.

As shown in the table above, all the improved recommendation algorithms perform better than Naive Bayes. Especially, results of M-SVD+SNA CF (1) and M-SVD+SNA CF (2) are better than M-SVD. Thus, it is necessary to add the citation information into recommendation system. In addition, M-SVD+SNA CF (2) shows better effect than M-SVD+SNA CF (1). Multiple features considered in the algorithm have a positive impact on the recommendation results.

V. CONCLUSIONS

With the accumulation of science citation information and the extraordinary development of recommendation algorithm, users are willing to have access to a better recommendation service on citation data set. In order to alleviate the problems of information overload in citation information, more and more improved algorithms are proposed to improve the recommendation system. This paper proposed a Collaborative filtering recommendation algorithm combined with citation data analysis in order to improve the accuracy of predicted results. The experiments have shown that the proposed algorithm is proved to be effective and the accuracy of recommendation can be improved.

ACKNOWLEDGMENT

This work is partially supported by China Postdoctoral Science Foundation (2014M561293), Science and Technology Planning Project of Jilin Province, China

(60171009201505200 64JH; 20130101179JC-03), Public computing platform in Jilin Province, China.

REFERENCES

[1] Kim, Meen Chul; Chen, Chaomei. A scientometric review of emerging trends and new developments in recommendation systems[J]. SCIENTOMETRICS. Volume: 104 Issue: 1 Pages: 239-263. Published: JUL 2015.

[2] . Champiri, Zohreh Dehghani; Shahamiri, Seyed Reza; Salim, Siti Salwah Binti. A systematic review of scholar context-aware recommender systems[J]. EXPERT SYSTEMS WITH APPLICATIONS. Volume: 42 Issue: 3 Pages: 1743-1758. Published: FEB 15 2015.

[3] Kenekayoro, Patrick; Buckley, Kevan; Thelwall, Mike. Hyperlinks as inter-university collaboration indicators[J]. JOURNAL OF INFORMATION SCIENCE. Volume: 40 Issue: 4 Pages: 514-522. Published: AUG 2014.

[4] Ozel, Bulent. Collaboration structure and knowledge diffusion in Turkish management academia[C]. Joint Meeting of the 7th International Conference on Webometrics, Informetrics and Scientometrics. SCIENTOMETRICS. Volume: 93 Issue: 1 Pages: 183-206. Published: OCT 2012.

[5] http://www.cdblp.cn[Z].

[6] http://academic.research.microsoft.com/About[Z].

[7] ManhCuongPhamAClusteringApproachforCollaborativeFilteringecommendation Using Social Network Analysis Journal of Universal Computer Science[J], vol. 17, no. 4 (2011), 583-604.

[8] Ji L. A collaborative recommendation method based on user network community with weighted spectral analysis[J]. Journal of Dalian University of Technology, 2010.

[9] Chen Y, Garcia E K, Gupta M R, et al. Similarity-based Classification: Concepts and Algorithms[J]. Journal of Machine Learning Research, 2008, 10(2):747-776.

[10] BleiDM,NgAY,JordanMI.Latentdirichletallocation[J].TheJournal ofMachineLearning Research,2011,3:993-1022.