Research on Personalized Recommendation Method of Academic Resources based on Hadoop

Mingxuan Caotian, Xueyan Chen*, Shuoxuan Fang and Dejun Chen
School of Information Engineering, Wuhan University of Technology, 430070
Wuhan, China

ABSTRACT

Facing the huge amount of academic resources, it is difficult for users to find the needed resources accurately through keyword search. On the basis of collaborative filtering, a personalized recommendation method for academic resources is proposed by fusing user attributes, academic resource attributes, and users' scores on academic resources, effectively alleviate the scarcity of scoring and cold start problems, improve the accuracy of the recommended method. Combined with the advantages of Hadoop distributed computing, design and implementation the highly effective academic resources personalized recommendation method which based Hadoop, the validity and of the design are proved by experiments.

INTRODUCTION

With the continuous development of science and technology, new technologies, new disciplines continue to emerge, the number of academic resources shows explosive growth. When users face a large amount of academic resources, often through the way of the keyword matching to retrieve the required resources, it can easily lead to the phenomenon that some academic resources which are valuable to users are not retrieved, thus reducing the utilization rate of academic resources.

Personalized recommendation is a very effective information filtering technology, with using data mining and analysis techniques, automatically, intelligently, and personalized initiative from the vast amounts of information for users to find and push to its content of interest, has important application value[1], this method is widely used in the field of e-commerce, social networking and other Internet. The recommended method is the core of the recommended system. The quality of the recommended method determines the accuracy and efficiency of the recommended system. The common recommendation algorithms include recommendation based on collaborative filtering, content-based recommendations, recommendation based on association rules, and recommendation based on graph architecture. Comprehensive comparison of the common recommended method, the recommendation based on collaborative filtering give consideration to the accuracy and efficiency of recommendations, it is the most widely used method, but the scoring sparse and cold start affected its initial recommendation method[2]. In order to alleviate the sparseness problem to some extent, Ying Y. proposed a hybrid collaborative filtering recommendation method based on project score prediction[3]. In order to solve the cold start problem, Peng S. proposed a method based on the Jaccard coefficient for the comprehensive similarity calculation[4]. Song S. proposed a collaborative filtering algorithm which based on Slope-one[5]. Most methods, however, are aimed at one problem only. In order to take into account the accuracy and efficiency of cooperative filtering algorithm, this paper combines the...
attributes of users, the attributes of academic resource with the attributes of user’s resource scores, and proposes a personalized recommendation method of academic resources. To some extent, the problems of sparseness and cold start existing in recommendation based on collaborative filtering are solved.

The personalized recommendation method of academic resources needs to dig out the academic resources which meet the user’s preference from the massive academic resources and take the initiative to the users. With the rapid increase in the number of academic resources, the traditional stand-alone centralized information processing architecture is limited. Hadoop is an open source, free distributed computing platform from the Apache Software Foundation (ASF), with high scalability, availability, reliability, and many other advantages[6]. As a distributed system infrastructure, Hadoop does not require developers to understand the underlying structure of the distributed, it can take full advantage of the functions of the cluster to achieve high-speed computing system and storage[7]. Therefore, this paper adopts Hadoop technology to realize the personalized recommendation of academic resources. This method can greatly improve the scalability, usability and reliability of the recommendation system, and greatly improve the efficiency of individual recommendation.

DESIGN OF PERSONALIZED RECOMMENDATION METHOD OF ACADEMIC RESOURCES

The personalized recommendation process for academic resources is shown in Figure 1.

![Figure 1. Academic resources personalized recommendation flow chart.](image)

(1) Attribute Similarity Matrix of User

Conventional attribute similarity calculation is generally used Jaccard coefficient. When calculating the similarity using the Jaccard coefficient, the weight of the attribute is not taken into account, and it is possible to reduce the similarity calculation accuracy by treating all the attributes equally. Academic resources sharing platform users to calculate the similarity of the attributes include the author age (segmentation), education (doctor, master, professor, etc.), research direction, obviously the importance of research direction is greater than age and education. Therefore, this paper presents a method of attribute information entropy as attribute weight, the attribute information entropy is calculated as shown in Equation 1:

\[
I(a_k) = -\sum_{i=1}^{n} p(v_i) \times \ln(p(v_i))
\]

(1)

Where \(I(a_k)\) represents the information entropy of user attribute \(a_k\), \(p(v_i)\) represents the probability that the \(a_k\) attribute value is \(v_i\), as shown in Equation 2:

\[
p(v_i) = \frac{N_v}{C_a}
\]

(2)

Where \(N_v\) is the number of users whose attribute \(a\) is \(v\) and \(C_a\) is the number of users with attribute \(a\). In the academic resource sharing platform, obviously \(C_a\) is the number of all users.

The similarity calculation for user attributes is shown in Equation 3:
\[ \text{sim}_{\text{attr}}(u,v) = \left( \sum_{k=1}^{n} I(a_k(x)) \times S(u,v,a_k) \right) / \left( \sum_{k=1}^{n} I(a_k) \right) \] (3)

Where \( S(u,v,a_k) \) indicates whether the attributes \( a_k \) of the user \( u,v \) are the same, the same take 1, different take 0.

(2) Synthetical Similarity of User

The user's similarity is weighted by the similarity of user attributes and user rating similarity. The concrete calculation is shown in Equation 4:

\[ \text{sim}_u(u,v) = \alpha \times \text{sim}_{\text{attr}}(u,v) + (1 - \alpha) \times \text{sim}_{\text{rate}}(u,v) \] (4)

\( \alpha \) is the adjustment factor of user attribute, \( \alpha \in [0,1] \).

(3) Pseudo Predictive Score Matrix

According to the user's academic resources, as well as the user's synthetical similarity, calculate the user's forecast score for unrated academic resources, and filled into the scoring matrix, thus to a certain extent weakened the sparsity of the scoring matrix. The specific calculation is shown in Equation 5:

\[ P_{a(u,v)} = \left( \sum_{i \in N} (\text{sim}_{ui}(u,v) \times R_{ui}) \right) / \left( \sum_{i \in N} \text{sim}_{ui}(u,v) \right) \] (5)

Where \( N \) indicates that take the first \( N \) most similar users to score and predict, and to fill scoring matrix.

(4) Similarity Matrix of Academic Resource

The similarity matrix of academic resources is consistent with the solution of user similarity matrix. This paper chooses the research direction attributes of the resource to calculate the similarity.

(5) Synthetical Similarity of Academic Resources

The synthetical similarity of academic resources is weighted by the similarity of academic resource attributes and the similarity of academic resource scores. The concrete calculation is shown in Equation 6:

\[ \text{sim}_{ij}(i,j) = \beta \times \text{sim}_{\text{attr}}(i,j) + (1 - \beta) \times \text{sim}_{\text{rate}}(i,j) \] (6)

\( \beta \) is the adjustment factor of academic resource attribute, \( \beta \in [0,1] \). At this point the similarity of academic resource scores is calculated from the post-filling scoring matrix. Compared to the original scoring matrix, sparseness is low.

(6) Forecast Score Results

For the user resource scoring matrix and the academic resource similarity matrix after filling, the average weighting strategy is used to calculate the predicted score of the remaining users' unrated academic resources. The specific calculation is:

\[ P_{a,i} = \bar{R} + \left( \sum_{j \in N} (\bar{R}_{ui} - \bar{R}_j) \times \text{sim}_{ij}(i,j) \right) / \left( \sum_{j \in N} \text{sim}_{ij}(i,j) \right) \] (7)

Evaluate the results of the forecast (the forecast plus the forecast in step 3) to make the recommendation to the user.

The personalized recommendation method reduces the score sparseness by pre-filling the user score matrix, thereby improve the recommended accuracy. By adjusting the similarity between attribute similarity and score similarity, the problem of cold start is alleviated and the accuracy of the recommendation system is further improved. Therefore, theoretically, the method is effective.

**DESIGN OF INDIVIDUALIZED RECOMMENDATION METHOD OF ACADEMIC RESOURCES BASED ON HADOOP**

The personalized recommendation method of academic resources firstly uses the collaborative filtering based on the user to fill the scoring matrix, and then uses the collaborative research based on the academic resources to predict the score,
although the accuracy of the recommended system is improved, it is at the expense of efficiency. With the growth of the number of shared platform user resources, the computing time spent by the personalized recommendation method will also be greatly increased. Therefore, a reasonable realization of the way to improve the efficiency of personalized recommendation is essential.

This section will be designed using Hadoop to implement object-based collaborative filtering recommendations and user-based collaborative filtering recommended methods. It should be noted that ASF’s open source project mahout also provides a collaborative filtering recommendation based on Hadoop, but with the algorithm designed in this paper is different, mainly reflects in: ① mahout package the collaborative filtering based on items, and can not make changes to the intermediate process, and this paper needs to integrate the similarity of item score and the similarity of item attribute. ② Mahout is based on HDFS as a data source, HDFS file storage rules are "write once, read many times", that is, the file can not be changed. In this paper, HBase is used as the data source due to the need to fill the user object score matrix. The process of collaborative recommendation based on Hadoop’s academic resources is shown in Figure 2.

As can be seen from Figure 2, the whole process consists of six MapReduce. In the figure, $u$ is the user, $i$ is the academic resource, $R_u$ is the user's score of the academic resource, and $R_{av}$ is the average score of the academic resource. $sim_i$ shows the similarity between $i$ and other resources. Take the pseudo code of the first MapReduce in Figure 2 as an example to illustrate the program implementation method, and other MapReduce implementations are similar. The symbol definitions in all the following pseudo-codes are the same as here.

**MR-I:** The average score information $<u, R_u>$ and $<i, u, R_u, R_{av}>$ of the user are output based on the score information $<u, i, R_u>$ of the input user's academic resource.

---

**The pseudo-code of MR-I map process.**

**Input:** $<u, i, R_u>$

**Output:** $i, <u, R_u>$

```java
String[] item_score=value.split;
Output.write(item_score[0], key+item_score[1]);
```

**The pseudo-code of MR-I reduce process.**

**Input:** $i, <u, R_u>$

**Output:** $i, R_u$ and $i, <u, R_u, R_{av}>$

for each $i, <u, R_u>$ DO

---

Figure 2. The flow chart of personalized recommendation method based on Hadoop.
THE VERIFICATION OF PERSONALIZED RECOMMENDATION METHOD OF ACADEMIC RESOURCES BASED ON HADOOP

The evaluation system of the recommended system mainly has the following three points [8]: (1) Mean Absolute Error (MAE); (2) Precision; (3) Recall rate.

The test sample set uses the Movielens data set, ml-1m, provided by the Grouplens team, which includes three files: movies.dat, users.dat, and ratings.dat representing the list of movies, the list of users, and the scoring list, the score list includes 1000209 ratings for 3900 films from 6040 users with the score sparseness of 4.3%, 80% of which was chosen as the training set and 20% as the test set. As the tests have a lot of records, the number of valid items recommended by the recommended system increases, and it is not obvious in the precision and recall rate. Therefore, this paper selected MAE as an evaluation index.

In this paper, we first test the effect of the adjustment factor on the MAE to determine the best value, and then compare the personalized recommendation method and the traditional cooperative filtering recommendation method to obtain the MAE and the time of consumption under different nearest neighbor N values, and prove the rationality of the design.

Take the value of N to 20, 40, 60 respectively, \( \beta = 0 \), adjust the weight factor \( \alpha \), MAE changes shown in Figure 3: when \( \alpha \) is 0.55, the MAE is the smallest, at this point the personalized recommendation method works best.

Take the value of N to 20, 40, 60 respectively, \( \alpha = 0.55 \), adjust the weight factor \( \beta \), MAE changes shown in Figure 4. It can be seen from the figure, when \( \beta \) takes 0.45, MAE is the smallest, then the personalized recommendation algorithm works best. Therefore, when \( \alpha=0.55 \, \beta=0.45 \) personalized recommendation algorithm MAE minimum, the recommended effect is the best.

![Figure 3: The picture of relationship between \( \alpha \) and MAE.](image-url)
The values of $\alpha=0.55$, $\beta=0.45$ and $N$ are taken as 20, 40, 60, 80, 120, in this paper, the personalized recommendation algorithm (3) the collaborative filtering algorithm based on the item (2) and the user-based collaborative filtering algorithm (1) are shown in Figure 5. It can be found from the figure that the MAE value of the personalized recommendation method is obviously smaller than that of the cooperative filtering recommendation, which improves the accuracy of the recommended system effectively.

Let $N = 120$, test the time spent in the stand-alone recommendation method and the personalized recommendation method in Hadoop mode, in order to make the comparison results more fair and accurate, this paper Hadoop using pseudo-distributed installation, that is, using a computer on the five process model distributed environment. The results are shown in Figure 6.

Obviously in the Hadoop mode personalized recommendation method greatly reduces the consumption time, improve the efficiency of the recommended system. Programs developed in the Hadoop pseudo-distributed model can be deployed without modification to clusters of thousands of machines, with unparalleled
scalability of traditional implementations, to create the conditions for digging out the resources that the user may prefer from the massive academic resources.

CONCLUSIONS

In this paper, users face the massive academic resources, the use of traditional keyword matching search method may cause the problem of low resource utilization, a variety of recommended methods are analyzed and compared, and a personalized recommendation method is proposed to integrate user attributes, academic resource attributes and user resource scores. To a certain extent, to overcome the cold start and sparseness issues of collaborative filtering recommended, improve the accuracy of the recommended system. Design and implement a personalized recommendation process based on Hadoop, through the example of data test comparison, proved that the recommended efficiency greatly improved.

REFERENCES