A Book Recommendation Method Considered Readers and Literatures for University Literature Service

Xiao CHEN¹,a

¹Library of Lanzhou Jiaotong University, Lanzhou, Gansu, China

a21319768@qq.com

Keywords: Data mining, Collaborative Filtering, Recommendation, Cluster.

Abstract. Because of the varieties of reader selection and literature categories, the rating matrix is often sparse and diffused excessively, that leads an inefficiency book recommendation. Aiming that, considered the implicit relation between reader group and literature classes, a collaborative cluster recommendation method has been constructed, the balance factor \( \alpha \) is defined, which is used to coordinate the weight of reader group and literature classes. Finally, through the test and validity evaluation, the cluster method mentioned in the paper can get a more productive recommendation than traditional reader cluster method.

Introduction

According to diversification construction for university library, library resource has increased rapidly, in which, paper literature, shared electronic literature, and network literature are all included. So how to make a smart recommendation service to satisfy readers’ (students or teachers) requirements, and help readers find out the required literature, book, video, article and so on, has become an import problem for library service. That is to say, Construction of library has come into service integration and personalized recommendation period[1,2]. The personalized recommendation is based on vast access log, utilize the correlative machine learning methods, and provide the personalized service to the reader based on their group characteristics[3,4]. Collaborative filtering recommendation is currently the most widely used recommended technology, although the user interest can be found, but by the sparseness and scalability of the data, it can be integrated together with cluster generally, in order to get a higher recommended accuracy[5].

Discussion of traditional cluster focus on the single relation of book and readers, it ignore the owner-member relationship between individual and group, at the same time, the rating matrix generated by traditional method is too sparse to get a better cluster effect. Based on above, a collaborative recommendation method has been discussed, in which the cluster concerned reader group and literature classes have been considered.

Collaborative Filtering Recommendation

Sun[6] has described the general procedure of collaborative filtering recommendation, rate matrix expression, neighbor selection, and recommendation result generation. The detailed description is as the following:

1. Rating matrix expression. Construct the rating matrix of user – item, for a user or item, it can be represented by a row vector or a column vector of the matrix.

2. Neighbor user or item selection. Through finding the users who have similar evaluation with the target user, calculating and ranking the user similarity according to the traditional similarity calculation method, the neighbor selection can be finished.

3. Recommendation result generation. When the similarity list is generated, the first N users with higher similarity are selected, and the score of unrated items is predicted according to the score of the similar user. The higher score is selected according to the prediction score.
Cluster Considered Reader and Literature Class

In this paper, clustering method is used to aggregate readers with similar attribute information into the same class, which can alleviate a large number of calculations when solving nearest neighbors.

**Clustering Process.** Based on the correlation between readers’ characteristics and the types of literatures, this paper designs a collaborative filtering literature recommendation process. The process is shown as follow. (1) Constructs rating matrix for Reader – Literature, then follow (2) and (3) to continue the process respectively. (2) Implement the reader clustering and then go to (4). (3) Implement the literature clustering, then go to (7). (4) In the clustering results, look for the class of the target reader, calculate its nearest neighbors, and go to (5). (5) Rate the unrated items and go to (6). (6) Form a reader recommendation list and go to (9). (7) For the target reader records, construct and find its nearest neighbor literatures, then go to (8). (8) Form a literature recommendation list and go to (9). (9) Implement the comprehensive evaluation based on the reader and literature recommendation lists. (10) Give out the conclusion.

**Rate Matrix Construction.** In the process of establishing the reader-literature rating matrix, if we select the literature number as a item, and use the borrowing status is as a rating score, a sparse 0-1 matrix can be got, for the recommendation, it is not significant. In fact, the borrowing or using status of the same kind of literatures also reflects the degree of concern of this literature [7]. This paper’s work is based on this, the borrowing books and viewing electronic source online were referred to as accessing literature collectively.

For the above scenario, a list of m readers and a list of n literature classes were created,

\[
\text{Reader} = \{r_1, r_2, \cdots, r_m\} \quad (1)
\]

\[
\text{Lit-Class} = \{c_1, c_2, \cdots, c_n\} \quad (2)
\]

Then the reader-literature class matrix \( RL_{m \times n} \) can be got, \( RL(i,j) \) is the number of accessing some literatures.

**Clustering Analysis.** After the reader-literature rating matrix is obtained, the readers and the literature class were clustered by row and column calculations respectively. The algorithm includes three parts, reader cluster, literature cluster and comprehensive recommendation.

The process of reader clustering Analysis is as the following. (1) Input the \( RL_{m \times n} \). (2) Carry out data preprocessing of Reader, and determine the appropriate number of clusters \( k \) and \( k \) cluster centers, which is expressed as \( CR=\{cr_1, cr_2, cr_3, \cdots, cr_k\} \), then use K-Means method to cluster the \( \text{Reader} \). (3) Forecast and rate the null value of \( RL_{m \times n} \) matrix. (4) Generate the reader nearest neighbor. (5) Generate similar sets of readers, and sort them from high to low according to their similarity.

The process of literature clustering Analysis is as the following. (1) Input the \( RL_{m \times n} \). (2) Carry out data preprocessing of \( \text{Lit-Class} \), and determine the appropriate number of clusters \( l \) and \( l \) cluster centers, which is expressed as \( CT=\{ct_1, ct_2, ct_3, \cdots, ct_l\} \), then use K-Means method to cluster the \( \text{Lit-Class} \). (3) Generate the literature nearest neighbor. (4) Generate similar sets of literatures, and sort them from high to low to finish Top-n nearest neighbor recommendation.

The final comprehensive recommendation need to consider both of Eq.(1) and (2), here Eq. (3) has been designed to finish the comprehensive optimization.

\[
\text{Similarity}(u,v) = \alpha \text{SimReader}(u,v) + (1-\alpha)\text{SimLit-Class}(l_u,l_v) \quad (3)
\]

\( \text{Similarity}(u,v) \) is used to express the comprehensive similarity of reader \( u \) and \( v \), \( \text{SimReader}(u,v) \) is the rating similarity of reader \( u \) and \( v \), \( \text{SimLit-Class}(l_u,l_v) \) is the class similarity of the literatures which reader \( u \) and \( v \) have accessed respectively. Here, Euclidean distance, Cosine and Pearson methods have been selected to finish the similar calculation, their description are shown as Eq. (4-6).

\[
d(u,v) = \sqrt{\left( \sum(u_i - v_i)^2 \right)^n} \quad (4)
\]
\[ Sim(u, v) = \cos(\vec{u}, \vec{v}) = \frac{\vec{u} \cdot \vec{v}}{|\vec{u}| |\vec{v}|} = \frac{\sum_{i \in G} R_{u,i} R_{v,i}}{\sqrt{\sum_{i \in G} R_{u,i}^2} \sqrt{\sum_{i \in G} R_{v,i}^2}} \] (5)

\[ sim(u, v) = \frac{\sum_{i \in I_u} (r_{u,i} - \bar{r}_u) \times (r_{v,i} - \bar{r}_v)}{\sqrt{\sum_{i \in I_u} (r_{u,i} - \bar{r}_u)^2} \sqrt{\sum_{i \in I_v} (r_{v,i} - \bar{r}_v)^2}} \] (6)

When \( Similarity(u, v) \) is calculated, \( G = m, u \) and \( v \) express reader \( u \) and \( v \), \( R \) is its properties. When \( SimLit Class(l_u, l_v) \) is calculated, \( G = n, u \) and \( v \) express literature \( u \) and \( v \), \( r \) is its rating score, \( \bar{r} \) is its mean value, \( \alpha \) is balance factor in Eq. (3) which should be confirmed via experiments testing.

The comprehensive recommendation process is defined as follow.

1. For the reader cluster, randomly selects \( k \) rows from all rows (columns) of the matrix, and generates \( CR \) as the initial reader clustering center. For the literature cluster, randomly selects \( l \) rows of the matrix, and generates \( CT \) as the initial literature clustering center.
2. Repeats steps (3) and (4) until the clustering center (\( CR \) and \( CT \)) does not change.
3. According to the mean value of each cluster object (center object), calculate the distance between each object and these central objects, and adjust the corresponding object division according to the minimum distance.
4. Recalculate the clustering centers of each cluster.

Test and Analysis

Data Preparation. The readers and literatures accessing data set of Library of Lanzhou Jiaotong University in 2014, was selected as experimental data set, it had 1000 accessing records and 50 reader records whose accessing times was more than 8 times, and the ratio of the training data set and test data set was 800:200.

First, the original data is transformed into a reader-literature class matrix by preprocessing such as cleaning, integration, and reduction. In this paper, based on Chinese Library Classification, we classify the column attributes into 22 categories as reader vector features, construct the matrix \( RL \). Matlab7.11b was selected as analysis tool.

Similarity Evaluation and Weight \( \alpha \) Determination. Euclidean distance, Cosine and Pearson were used as the similarity evaluation criteria, in Top-N method, the clustering values \( k \) and \( l \) are both 20, and the number of neighbors is increased from 5 to 100, increment is 5. The distribution of similarity is shown as Fig. 1. The similarity by \( Euclidean distance \) and \( Cosine \) method is relative concentrated, its classifying quality is worse than \( Pearson \).

Figure 1. Similarity Distribution of Three Methods.
For the determination of weight $\alpha$, by taking different values of $\alpha$, the mean absolute error (MAE) can be calculated according to Eq. (7). The value of $\alpha$ can be determined by observing MAE.

$$MAE = \frac{\sum_{i=1}^{N}|p_i - q_i|}{N}$$

(7)

$p_i$ is the predicted value of item, $q_i$ is the actual value of item, $N$ is the total quantity of items. The value changing is described in Fig. 2. With the increase of $\alpha$, the MAE value decreased firstly and then increased, and when $\alpha = 0.6$, the MAE was the lowest, so the value of $\alpha$ is 0.6.

![Figure 2. Values of MAE according to $\alpha$ Changing.](image)

Algorithm Performance Evaluation. The Precision, Recall and F-measure [6] was as the evaluation indicators, taking Pearson method as an example., their classification error has been described in Fig. 3.

$$Precision = \frac{N(Right)}{N(All)}$$

(8)

$$Recall = \frac{N(Right)}{U(All)}$$

(9)

$$F\text{-meas}ure = \frac{2 \times Precision \times Recall}{Precision \times Recall}$$

(10)

![Figure 3. Changing of the Precision, Recall and F-measure.](image)
Figure 4. Comparison of the MAE Value.

With the increase of the number of neighbors, the recommendation precision will decrease. The recall will increase with the increase of Top-N, and the F-measure value does not change much.

The method mentioned in this paper is compared with the traditional collaborative recommendation method based on reader clustering, the value changing of MAE is described in Fig. 4 with using Pearson method.

For the recommendation algorithm mentioned in this paper, the MAE value has a rapid decline trend at the beginning. When the number of neighbors is 30, the MAE reaches a minimum value, then with the increase, the MAE becomes bigger, because the number of neighbors becomes bigger and bigger, so that the nearest neighbor readers cannot be accurately selected. When the number of neighbors exceeds 45, the algorithm performance is close to the original reader-based algorithm.

**Algorithm Effectiveness Evaluation.** In order to verify the effectiveness of this mentioned method, we selected 120 students of four classes, and recommended 10 books to the target students from 200 new books. The effectiveness evaluation is based on $P$(Precision) and $F$(Fallout), their description is as Eq. (11), $N(success)$ is the number of successful recommendation, $N(recommend)$ is the number of recommendation, $N(failure)$ is the number of failure recommendation, $N(all)$ is the total number of books. Table 1 is the final result of recommendation, $P_0$ and $F_0$ are the precision and fallout of the original recommendation, $P$ and $F$ are based on method mentioned in this paper.

$$P = \frac{N(success)}{N(recommend)}, F = \frac{N(failure)}{N(all)}$$  \hspace{1cm} (11)

<table>
<thead>
<tr>
<th>Number of readers</th>
<th>Number of books</th>
<th>$P_0$</th>
<th>$P$</th>
<th>$F_0$</th>
<th>$F$</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>50</td>
<td>54</td>
<td>59</td>
<td>23.1</td>
<td>21.7</td>
</tr>
<tr>
<td>25</td>
<td>100</td>
<td>61</td>
<td>68</td>
<td>21.3</td>
<td>18.5</td>
</tr>
<tr>
<td>35</td>
<td>150</td>
<td>69</td>
<td>77</td>
<td>18.2</td>
<td>15.7</td>
</tr>
<tr>
<td>45</td>
<td>200</td>
<td>75</td>
<td>78</td>
<td>16.1</td>
<td>14.6</td>
</tr>
</tbody>
</table>

Table 1. The Result of Recommendation.

With the number increasing of readers and books in the beginning period, the recommended success rate shows an upward trend, the error rate showed a downward trend, the mentioned method show a better recommended ability than the reader clustering method. When the number of reader increase to a certain extent, the recommendation ability of the mentioned method is close to the traditional method, but still has certain advantages.

**Conclusion**

Based on the accessing records of readers and literatures, a collaborative recommendation method has been designed, which has considered the relationship between readers and literature class, and implemented synchronous cluster for readers and literatures. The experiment shows that this method
has a better recommended ability than traditional reader clustering method, especially for university library users. In addition, if the number of accessing records increase significantly, this method can improve the time efficiency.

References


