Multidimensional Expert Scoring Model of Institutional Repository Based on PCA

Shi-yong XIONG¹, Liang CHANG¹,* , Ai-rong XIE² and Chen LU¹

¹Chongqing University of Posts and Telecommunications, Chongqing, China
²Communication NCO Academy Army Engineering University of PLA, Chongqing, China

*Corresponding author

Keywords: PCA, Institutional repository, Multidimensional, Expert scoring.

Abstract. Generally, expert information scoring has many dimensions in institutional repository. In order to reduce the computational time and improve the computational efficiency in the process of expert recommendation based on institutional repository, a multidimensional expert scoring model of institutional repository based on PCA is proposed. Firstly, the multidimensional expert scoring information matrix of institutional repository is constructed according to the feature dimension of the expert in the institutional repository. Then, using PCA algorithm to calculate the dimensionality reduction of multidimensional expert scoring information matrix of institutional repository. Finally, the multidimensional expert scoring model of institutional repository is obtained according to the 95% component information of the original space. Comparing the reconstruction errors between expert scoring test samples and non-expert scoring test samples, it shows that the model can effectively represent multidimensional expert scoring information of institutional repository.

Introduction

With the continuous development of Internet technology and the increasing of information resources. It is necessary to build the institutional repository, in order to share and integrate information resources. However, at this stage, the institutional repository contains not only scientific research data, but also scientific research experts. Moreover, most of the institutional repositories provide services for the traditional academic information resources such as papers, journals, conferences, reports, etc. But, experts who are carriers of tacit knowledge rarely provide recommendation services. In order to improve the accuracy and speed of recommendation, a multidimensional expert scoring model of institutional repository based on PCA is proposed for expert recommendation, because there are many dimensions of expert information scoring in institutional repository.

Related Research

At present, the recommendation of personalized information in institutional repository is relatively few [1]. And the current recommendation methods are similar article recommendation, popular browsing recommendation and the recommendation of the home page of institutional repository according to the upload time and downloads visits. The expert recommendation using institutional repository is rare, which makes the research on expert scoring model of institutional repository almost absent.

Up till now, the evaluation research of experts at home and abroad mainly focuses on the performance evaluation index system of scientific and technological experts. There is little research on expert evaluation in institutional repositories. Literature [2],[3] puts forward the mathematical model from the basic situation indicators of experts, training indicators, the hit rate of review projects, the dispersion rate of project scores and the rate of project review results. Literature [4] used the analytic hierarchy process to classify the expert’s index system into educational level, professional title, unit where the expert is located, scientific research achievements and awards, publication of papers and books, familiarity of the expert review field. And the weight calculation is
carried out. Literature [5] takes expert review experience, score deviation, review hit rate and review validity as evaluation indicators. Literature [6] uses expert papers, research projects, review performance and other evaluation indicators to conduct evaluation research. Literature [7] established the performance evaluation indicators of experts in scientific research and evaluation of work performance, and established a comprehensive evaluation mathematical model. Literature [8] used the analytic hierarchy process to establish an expert evaluation system for experts’ research topics, books, awards, professional title, etc. Using regression analysis to determine indicators weights, and proposed expert scoring mathematical model. Literature [9] established a review expert counter-evaluation index system by using the experts’ personal qualifications, papers, social rewards, dispersion rate and other indicators. Literature [10] selected 24 evaluation indicators from the relevant literature, and then 11 indicators such as scientific research achievements, review hit rate, review success rate, review reputation, etc. were determined by sampling survey, interview, and consulting expert. The weight of the indicators was determined by the entropy weight method, and the selection model was constructed. Literature [11] established a scientific and technological entries model based on tree graph, according to the characteristics of scientific and technological entries. And an evaluation method of scientific and technological experts based on the model of scientific and technological entries is given.

Multidimensional Expert Scoring Model of Institutional Repository Based on PCA

Statement of Problem

In order to better communicate and utilize the experts of the tacit knowledge carrier in the institutional repository, it is necessary to recommend the experts in the institutional repository. A multidimensional expert recommendation method based on institutional repository is proposed.

To conduct multidimensional expert recommendation based on institutional repository, the first thing is to establish the multidimensional expert scoring model of institutional repository. The main dimensions of expert characteristics in institutional repository include: age, professional title, educational background, work unit, scientific research projects, academic papers, awards, patents, writings, research reports, personnel training, guidance papers, etc.

Professional title can be divided into: assistant, lecturer, associate professor (associate senior engineer), professor (senior engineer).

Educational background can be divided into: undergraduate degree, master's degree, doctoral degree and postdoctoral degree.

Work units can be divided into: subordinate colleges and universities, provincial colleges and universities, municipal colleges and universities, private colleges and independent colleges.

Scientific research projects can be divided into: national scientific research projects, provincial and ministerial scientific research projects, municipal scientific research projects and other types of scientific research projects.

Academic papers can be divided into: SCI papers, EI papers, ISTP papers and other journal papers.

Awards can be divided into: national awards, provincial and ministerial awards, municipal awards and other awards.

Patents can be divided into: invention patents, utility model patents, design patents and software works.

Writings can be divided into: monographs, editions, translations and other writings.

Research reports can be divided into: international academic research reports, major domestic academic research reports, domestic general academic research reports, and other research reports.

Personnel training can be divided into: the number of postgraduates, the duration of tutors for postgraduates, the number of doctoral students, the duration of tutors for doctoral students.

Guidance papers can be divided into: undergraduate papers, postgraduate papers, doctoral papers, and other guidance papers.

In the process of expert recommendation, because of calculating so many expert information
scoring dimensions, it needs a lot of resources, and the computational efficiency is low, which is insufficient to meet actual needs. Therefore, it is necessary to reduce the expert information scoring dimension, retain the main information, improve the computational efficiency and reduce computational time to meet the actual needs.

**Model Construction**

In order to solve the problem of too many expert information scoring dimensions in the institutional repository, a multidimensional expert scoring model of institutional repository is established by using PCA method \[12\].

According to the characteristics of experts in the institutional repository, the multidimensional expert scoring information matrix of institutional repository is constructed:

\[
M = [M_1, M_2, \ldots, M_k]^T
\]  

(1)

PCA is used to reduce the dimensionality of the multidimensional expert scoring information matrix. The derivation process is as follows:

\[
M = \begin{bmatrix}
m_{11} & m_{12} & \cdots & m_{1n} \\
m_{21} & m_{22} & \cdots & m_{2n} \\
\vdots  & \vdots & \ddots & \vdots \\
m_{k1} & m_{k2} & \cdots & m_{kn}
\end{bmatrix}
\]  

(2)

Calculate the covariance matrix of matrix M according to formula (2):

\[
\sum_{k=1}^{1} \sum_{i=1}^{k} (m_{ij} - \overline{m}_j)(m_{ij} - \overline{m}_j) \\
\sum_{k=1}^{2} \sum_{i=1}^{k} (m_{ij} - \overline{m}_j)(m_{ij} - \overline{m}_j) \\
\vdots \\
\sum_{k=1}^{n} \sum_{i=1}^{k} (m_{ij} - \overline{m}_j)(m_{ij} - \overline{m}_j)
\]  

(3)

\[
\frac{1}{k-1} \sum_{i=1}^{k} (M_i - \overline{M})(M_i - \overline{M})^T
\]  

(4)

In formula (4), \(k\) is the number of samples, \(\overline{M}\) is the sample mean, and is a column vector, \(M_i\) for the \(i\) sample, and is also a column vector.

The eigenvalues and eigenvectors of the covariance matrix of the matrix \(M\) are calculated by singular value decomposition:

\[
SVD(\sum) = [U, S, V]
\]  

(5)

The matrix \(U\) and the matrix \(S\) are as shown in the formula (6) and the formula (7).

\[
U = [u_1, u_2, \ldots, u_k, \ldots, u_n]
\]  

(6)

\[
S = \begin{bmatrix}
S_{11} & 0 & \cdots & 0 \\
0 & S_{22} & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & S_{nn}
\end{bmatrix}
\]  

(7)

In formula (5), \(U\) is all the eigenvectors of the covariance matrix. The eigenvectors are sorted from large to small according to the size of eigenvalues. The dimension of \(U\) is \(n \times n\), \(U\) is called the dimensionality reduction matrix. \(S\) is the eigenvalues of the covariance matrix, and \(S\)
is a diagonal matrix, the dimension of $S$ is $n^*n$.

Selecting the first $k$ column of $U$ as the dimensionality reduction matrix $U_k$ whose dimension is $n^*k$, which is used to reduce the sample to $k$ dimensional.

**Selection of $k$ value in dimensionality reduction**

The selection of the dimension $k$ determines the number of the eigenvectors in the dimensionality reduction matrix $U$ [13]. The larger the value of $k$, the more eigenvectors in the dimensionality reduction matrix $U$ are selected, the smaller the error of dimensionality reduction, and the more original characteristics can be retained, and vice versa.

Matrix $S$ is used here to calculate (the 95% component information of the original space is selected):

$$0.95 \leq \frac{\sum_{a=1}^{k} S_{aa}}{\sum_{a=1}^{n} S_{aa}}$$

(8)

Obtain the multidimensional expert scoring model $W$ of institutional repository after PCA dimensionality reduction. The dimension $W$ is $n^*k$:

$$W = M \cdot U_k$$

(9)

**Experimental Results and Analysis**

**Evaluation Index**

In this experiment, using the sample reconstruction to evaluate the experiment. Reconstruct the test sample and compare it with the original sample [14]. Using formula (10) to calculate sample reconstruction error, $X'$ is the data from sample reconstruction, as shown in formula (11).

$$error = \frac{1}{m} \left\| X - X' \right\|_F$$

(10)

$$X' = WW^T X$$

(11)

**Analysis of Experimental Results**

This research relies on the key projects at provincial and ministerial level in Chongqing (“Chongqing Federation of Social Science Circles Institutional Repository Data Platform System R&D and Application”, 2016WT04). The experimental data set comes from the Chongqing Federation of Social Science Circles. The data set contains scores for at least 36 characteristic dimensions for each expert, ranging from 0 to 100, and each characteristic dimension is scored precisely to the last two decimal points. The data set used in this experiment included 128 experts’ characteristic dimension scores.

This experiment performs PCA dimensionality reduction processing, and draws a histogram of variance contribution rate as shown in Figure 1.

From the histogram of variance contribution rate, we can see that the first principal component explains less than 35% variance, and the first six principal components can reach the variance contribution rate of more than 95%.

The percentage of variance (POV) and the percentage of cumulative variance (POCV) of the first 10 principal components are calculated as shown in Table 1.
Table 1. POV and POCV of principal component.

<table>
<thead>
<tr>
<th>Principal Components</th>
<th>POV (%)</th>
<th>POCV (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>33.92</td>
<td>33.92</td>
</tr>
<tr>
<td>2</td>
<td>19.17</td>
<td>53.09</td>
</tr>
<tr>
<td>3</td>
<td>15.12</td>
<td>68.21</td>
</tr>
<tr>
<td>4</td>
<td>13.00</td>
<td>81.21</td>
</tr>
<tr>
<td>5</td>
<td>11.49</td>
<td>92.70</td>
</tr>
<tr>
<td>6</td>
<td>2.74</td>
<td>95.44</td>
</tr>
<tr>
<td>7</td>
<td>1.22</td>
<td>96.66</td>
</tr>
<tr>
<td>8</td>
<td>0.71</td>
<td>97.37</td>
</tr>
<tr>
<td>9</td>
<td>0.47</td>
<td>97.84</td>
</tr>
<tr>
<td>10</td>
<td>0.33</td>
<td>98.17</td>
</tr>
</tbody>
</table>

Figure 1. Histogram of Variance Contribution Rate.

As can be seen from the table 1, the first six principal components cover the variance of 95.44% of the data, and the first 10 principal components cover 98.17% of the variance. This show that the number of principal components that need to be retained can be determined by the eigenvalue analysis. And under the condition that the overall information loss of the data set is small, the data characteristic dimension can be greatly reduced.

The reconstructed error experimental analysis is performed on the obtained dimensionality reduction matrix. Five sets of expert scoring test sample data and five sets of non-expert scoring test sample data were used for data reconstruction. The reconstruction error results of the two sets of data are shown in Table 2 and Table 3, respectively.

Table 2. Reconstruction errors of expert scoring test samples.

<table>
<thead>
<tr>
<th>Test Samples</th>
<th>Reconstruction Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.2975e+003</td>
</tr>
<tr>
<td>2</td>
<td>5.2503e+003</td>
</tr>
<tr>
<td>3</td>
<td>5.0787e+003</td>
</tr>
<tr>
<td>4</td>
<td>4.0378e+003</td>
</tr>
<tr>
<td>5</td>
<td>4.1688e+003</td>
</tr>
</tbody>
</table>
Table 3. Reconstruction errors of non-expert scoring test samples.

<table>
<thead>
<tr>
<th>Test Samples</th>
<th>Reconstruction Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>337.9450e+003</td>
</tr>
<tr>
<td>2</td>
<td>295.7592e+003</td>
</tr>
<tr>
<td>3</td>
<td>487.2548e+003</td>
</tr>
<tr>
<td>4</td>
<td>154.6092e+003</td>
</tr>
<tr>
<td>5</td>
<td>372.4257e+003</td>
</tr>
</tbody>
</table>

By comparing the reconstruction errors of the two sets of data in Table 2 and Table 3, it can be seen that the reconstruction errors of expert scoring test samples are significantly smaller than those of non-expert scoring test samples. It shows that the dimensionality reduction matrix obtained can well represent the multidimensional expert information of the institutional repository.

Summary

Through the experimental results and analysis, it can be found that using PCA to reduce the dimension of expert scoring information matrix in the institutional repository can be effectively reduce the scoring dimension of expert and retain enough original data information, under the condition of retaining 95% of the original data. Therefore, the multidimensional expert scoring model of institutional repository based on PCA can effectively reduce the data dimension, and retain enough data information. It provides the basis for the follow-up multidimensional expert recommendation method based on institutional repository.

Acknowledgement

This research was financially supported by the Chongqing Federation of Social Science Circles.

References


