Research on resource construction and allocation model of university library based on principal component analysis

Runzhi Liu¹, Mei Li², Linlin Zhao³, Zhongmin Lv⁴ and Yu Zhang⁵

¹Beijing Sport University (China), Beijing, China
²Beijing Institute of Fashion Technology, Beijing, China
³Beijing Sport University (China), Beijing, China
⁴Beijing Sport University (China), Beijing, China
⁵Beijing Sport University (China), Beijing, China

Corresponding author and e-mail: Mei Li, victoriagood2001@sina.com

Abstract. University library is the document information center of a university. It is an academic institution serving for teaching and scientific research. The quality of document information resources construction directly affects the service level of library. In order to realize the objectivity, significances and operability of the resource construction and allocation of university library, this paper considers many factors that affect the service quality of the library and establishes the corresponding research model. In the model design, taking the general library of Peking University as the research object, the principal component analysis method is used to analyze and process the data through software, and the principal component is selected as the independent variable to establish the linear regression mode. The validity of the model is verified by comparing the predicted result with the actual value, and some reasonable suggestions are put forward for the resource construction and allocation of university library.

1. Introduction

As the literature and information center of teaching and scientific research services, University Library meets the development needs of disciplines and specialties in Colleges and universities and is an important part of the realization of educational tasks. The quality of library resources construction directly affects the ability and level of library service. Therefore, the library should adjust the construction and distribution of resources in time with the development of the school, so as to meet the needs of the main subject literature to the maximum extent, and at the same time, it can leave valuable books for the future.

The construction of university library resources is to meet the daily learning needs of college students to the greatest extent and at the same time, the library literature resources of the Embassy can give full play to its role. In the case of limited funds, it is necessary to allocate resources reasonably, collect the literature needed by readers, optimize the collection structure, improve the quality of literature, and provide better service and support for readers. In 2012, Jingliang Liu and Pei Li, taking Heilongjiang University Library as an example, proposed the method of combining "Delphi method, analytic hierarchy process (AHP) and formula" to determine the evaluation, index weight and construct the allocation model, and then scientifically and reasonably allocate the discipline resources.
allocation mechanism in Colleges and Universities, and give practical demonstration and Discussion on the construction of the mode [1]. In 2014, Xinchan Duan took the needs of readers as the starting point, and made corresponding suggestions on the construction of Library Literature Resources [2]. Shuping Zhao, Chunxuan Li and Xingye Zhang conducted exploratory research on the paper book construction of University Library in 2018, formulated the book procurement and distribution principles and divided the collection amount of various books according to the book construction level, which were 10% of the basic level, 40% of the learning level, 40% of the research level and 10% of the characteristic level [3]. In 2020, Xiaoxia Li applied dart model in the evaluation of digital library resource construction, distributed 300 questionnaires to users to collect data and according to the user's perspective, according to the final user satisfaction score to improve the service level of the library [4]. However, at present, the university library is developing towards a diversified direction and only providing the most basic literature resources can no longer meet the daily learning needs of students. Based on the principal component analysis method, this paper will add a variety of influencing factors to open up a new resource plate and re-carry out the construction and distribution of library resources.

2. Principle and algorithm
When using statistical analysis method to study multivariable problems, people want to get more information from fewer variables. However, in reality, the research object is often affected by multiple variables. When the information reflected by two factors overlaps, we call it the variable with certain correlation. Principal component analysis (PCA) is to delete the more relevant variables from all the variables originally proposed, and then construct relatively fewer new variables, so that these variables are unrelated to each other, and the information reflected by the original variables can be retained as much as possible in the new variables.

2.1. Basic principle of principal component analysis
Principal component analysis (PCA) is a statistical method to reduce the dimension. Through the orthogonal transformation of matrix, the variables with correlation are transformed into new random linear independent variables. In algebra, the covariance matrix of the original random variable is transformed into a diagonal matrix, and in geometry, the original coordinate system is transformed into a new orthogonal coordinate system, and the sample points are scattered in the most open P orthogonal directions. Then, the multidimensional variable system is reduced in dimension so that it can be transformed into a low dimensional variable system with high accuracy. Then, an appropriate value function is constructed to further transform the low dimensional system into a pile of systems.

2.2. Linear regression model based on principal component analysis
Regression analysis is an analysis of the quantitative relationship of objective things. It is a common method of mathematical statistics and a method to deal with the relationship between multiple variables. After the dimensionality reduction of principal component analysis, the principal component elements can be used as independent variables of linear regression model, and the predicted as dependent variables can be constructed to make regression prediction analysis on the research object.

The multiple regression prediction model of the number of Library visitors in the whole year is established as follows:

\[ Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \mu \]

Among them, \( Y \) is the annual number of visitors; \( X_1, X_2, \ldots, X_p \) is the influencing factor of each principal component; \( \beta_0, \beta_1, \ldots, \beta_p \) is the regression coefficient and \( \mu \) is the random error.

2.3. Computational procedure
Standardized acquisition of original index data p-dimensional random vector
\[ x = (X_1, X_2, \ldots, X_p)^T, \] N samples \( x_i = (X_{i1}, X_{i2}, \ldots, X_{ip})^T, i = 1, 2, \ldots, n, n > p \), the sample array is constructed and standardized
\[ Z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}, i = 1, 2, \ldots, n; j = 1, 2, \ldots, p \]

Where \( \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_{ij} \), \( s_j^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2 \), the normalized matrix \( Z \) is obtained.

Finding the correlation coefficient matrix of normalized matrix \( Z \)
\[ R = \begin{bmatrix} r_{ij} \end{bmatrix}_p \]
\[ r_{ij} = \frac{Z_{ij} \cdot Z_{kj}}{n-1} \]

Where \( \sum_{j=1}^{m} \lambda_j \geq 0.85 \) to determine the value of \( m \), the utilization rate of information reflected by variables can reach more than 85%. For each eigenvalue \( \lambda \), the unit eigenvector \( b_j^0 \) is obtained by solving equation \( Rb = \lambda b \).

The standardized vector index is transformed into principal component:
\[ U_j = z_i^T b_j^0, j = 1, 2, \ldots, m \]

Among them, \( U_1 \) is the first principal component, \( U_2 \) is the second principal component and \( U_m \) is the third principal component.

The M principal components were evaluated comprehensively
The final evaluation value is obtained by weighted summation of M principal components.

3. Experimental results and discussion
This paper collects the annual work report data of Peking University Library from 2005 to 2014. The annual number of Library entrants is used as the prediction value to test the feasibility of the regression model. The number of books collected, the number of e-books, the amount of database, the number of books lent, the proportion of funds used for the library, the proportion of funds used for daily maintenance, and the proportion of funds used for assets and equipment are taken as multiple variables. Using software for data processing and multiple regression analysis, the output results are as follows.

| Books (pieces) | E-book (piece) | Database | Number of copies on loan | Proportion of collection development funds (%) | Proportion of assets and equipment funds (%) | Proportion of daily maintenance funds (%) | Number of visitors |
|---------------|---------------|----------|-------------------------|-----------------------------------------------|---------------------------------------------|-------------------------------------------|-------------------|
| 2005          | 2295633       | 93000    | 373                     | 1011486                                       | 77                                          | 8                                         | 15                | 1695312          |
| 2006          | 3290000       | 433251   | 451                     | 1069049                                       | 80.33                                       | 6.83                                      | 12.84             | 2286294          |
| 2007          | 4890000       | 487148   | 460                     | 944625                                        | 85                                          | 3.8                                       | 11.2              | 2246006          |
Table 2. Regression model statistics.

|                        |            |
|------------------------|------------|
| Multiple R             | 0.995922095|
| R Square               | 0.99186082 |
| Adjusted R Square      | 0.963373691|
| Standard error         | 0.055010588|
| Observations           | 10         |

3.1. $R^2$ test

$R$ is the determination coefficient of regression analysis. It can be seen from table 2 that the change percentage of dependent variable explained by the fitting model is 99.1862282%, and the adjusted R square value reaches 0.963373691, indicating that the regression model has a high degree of fitting.

Table 3. Analysis of variance of regression model.

|                          | df (free degree) | SS (Sum of squares) | MS (Mean square error) | F         | Significance F |
|--------------------------|------------------|---------------------|------------------------|-----------|----------------|
| regression analysis      | 7                | 0.737552017         | 0.105364574            | 34.81786  | 0.02819848     |
| residual                 | 2                | 0.00605233          | 0.003026165            |           |                |
| total                    | 9                | 0.743604346         |                        |           |                |

3.2. $F$ test

Table 3 shows the test results of regression model. $F$ value is the significance coefficient of regression relationship, and is the actual significance probability of $F$ value, i.e. $p$ value. It can be seen from table 3 that 1, so F test is passed, there is a linear relationship between Y and X, and the regression equation has good fitting.

Table 4. Principal component regression coefficient and significance analysis table.

|                          | Coefficients     | Standard error  | t Stat         | P-value   |
|--------------------------|------------------|-----------------|----------------|-----------|
| Intercept                | 620.9644886      | 137.801624      | 4.506220394    | 0.045884  |
| X Variable 1             | 0.353374939      | 0.240250529     | 1.47086019     | 0.279149  |
| X Variable 2             | 0.042179666      | 0.246755722     | 0.170936933    | 0.88003   |
| X Variable 3             | 1.030550562      | 0.334697417     | 3.079064877    | 0.091268  |
| X Variable 4             | 0.858953531      | 0.170227712     | 5.045908936    | 0.037103  |
| X Variable 5             | -615.9210954     | 136.5795436     | -4.509614539   | 0.045819  |
| X Variable 6             | -350.7483697     | 77.73747494     | -4.5119599     | 0.045775  |
3.3. *t* test

From the results of regression coefficient analysis table 4 can be obtained: $t(\hat{\beta}_2)=0.042179666$, $t(\hat{\beta}_3)=1.03055062$, $t(\hat{\beta}_4)=0.858953531$, $t(\hat{\beta}_5)=-615.9210954$, $t(\hat{\beta}_6)=-350.7483697$, $t(\hat{\beta}_7)=-280.5827142$.

When the significance level is $\alpha=0.05$, the *t* distribution table shows that the critical value of *t* is 4.303.

It can be obtained that $|t(\hat{\beta}_1)|, |t(\hat{\beta}_2)|, |t(\hat{\beta}_3)|, |t(\hat{\beta}_4)|$ is greater than the critical value $t$, $X_4, X_5, X_6, X_7$ has passed the significance test, which shows that it has a significant impact on the number of visitors in the whole year.

Based on this, the regression prediction equation can be obtained as follows:

$$\hat{y} = 620.96 + 0.35X_1 + 0.04X_2 + 1.03X_3 + 0.86X_4 - 615.92X_5 - 350.75X_6 - 280.58X_7$$

3.4. Back substitution test

The predicted results are shown in Table 5:

| Observations | Forecast Y | Residual | Standardized residual |
|--------------|------------|----------|-----------------------|
| 1            | 0.00309136 | -0.00309136 | -0.119209158          |
| 2            | 0.919503414| 0.001910524 | 0.073673714           |
| 3            | 0.881404924| -0.022804954| -0.879405497          |
| 4            | 0.798500479| 0.042433062 | 1.636305327           |
| 5            | 0.675397146| -0.009501725| -0.366405889          |
| 6            | 0.959391101| -0.030825463| -1.188692678          |
| 7            | 0.879429521| 0.039361972 | 1.517877824           |
| 8            | 0.694545824| 0.013347706 | 0.514714753           |
| 9            | 1          | -2.84217E-14| -1.096E-12            |
| 10           | 0.75982291 | -0.030829761| -1.188858395          |

In order to verify the effectiveness of the above model, the actual value of the number of Library visitors from 2005 to 2014 was compared with the predicted value, and the data obtained is shown in Figure 1.
Figure 1. Comparison of predicted and actual number of visitors (after standardization) from 2005 to 2014.

Through comparison, the maximum relative error between the actual value and the predicted value of the number of Library visitors from 2005 to 2014 is not more than 0.039. Therefore, the multiple regression model has a strong prediction ability for the number of Library entrants in the whole year.

There are many links in the construction of document information resources, which is not a simple purchase of documents. How to do well in the construction of document information resources is a problem often discussed by the library circle. As the teaching content and teaching needs of colleges and universities are constantly changing, the library, as the literature information center of colleges and universities, should timely adjust the construction of literature resources according to the needs of readers, and provide strong support and services. The annual growth of books in Peking University Library is very large, with an increase of over 2.3 million in the past ten years, but there is no obvious increase in the attraction of users. Digital resources have brought the extension of library space. Digital library is the inevitable trend of library development in the future. Digital library breaks the space-time limitation of traditional library, and the cost of resource acquisition is low. Users can obtain the required resources in a short period of time through digital resources, and the digital platform can also collect the evaluation of the convenience and effectiveness of digital resources given by users after meeting their own information needs. In the process of resource construction, digital library needs to clarify user information behavior from the perspective of user demand and finally realize the efficient construction of digital resources.

In addition to the steady growth of paper books and the development of electronic resources, other functional services should be provided. With the increasing pressure of social work, further study has become a choice for many college students. University library has also become an important learning place for college students. Providing services such as electronic reading room, self-study room and book storage space is also the development direction of the library to enrich its own functions. It also needs to provide timely according to the needs of users, but it is followed by daily maintenance the maintenance work, the replacement of facilities, and the resource distribution structure of the library also need to be adjusted timely. In short, university library should not only give full play to the functional advantages of the traditional library, but also keep pace with the times and scientifically and reasonably carry out resource construction and allocation.
4. Conclusions
As the platform of knowledge dissemination, the resource construction and distribution of university library is a very important work. Based on the data of the number of entrants and their seven influencing factors in Peking University Library from 2005 to 2014, this paper uses the multiple regression model based on principal component analysis to predict the number of Library entrants in each year and compares with the actual value. The results show that the overall deviation of the prediction results is small, which provides a theoretical reference for the construction and allocation of library resources. However, the results of principal component analysis are still subject to the objectivity of indicators and the quality of data. Further research and improvement are needed in the selection of prediction impact indicators.

References
[1] Liu Jingliang, Pei Li. The Research on Fund Allocation Mode for Literature Resources Based on Subject Construction of Universities, A Case Study of Heilongjiang University of Chinese Medicine [J]. Heilongjiang University of Chinese Medicine Library. 2012, 6.

[2] Duan xinchan. On the Construction of document Resources in university library [J]. Management & Technology of SME. 2014, 11: 133-134.

[3] Zhao Shuping, Li Chunxuan, Zhang Xingye. Exploration on the Construction and Distribution principles of paper book Resources in university libraries [J]. Contemporary Education Research and Teaching Practice. 2018, 4.

[4] Li Xiaoxia. The Application of DART Model in the Evaluation of digital library Resources Construction [J]. The Library Journal of Henan. 2020, 1: 112-115.