Application of Clustering Analysis Algorithm in Digital Library

Yan Ge

Library of Jiangnan University, Wuhan, Hubei, 430056, China

geyan_sn@sina.com

Keywords: Clustering Analysis; K-means; Algorithm; Digital Library; Resource Analysis

Abstract. Clustering analysis embarks from the data itself to dig out the internal feature relations of the data and present the category information the data contains. Of the clustering methods, hierarchical clustering method, K-Means method, etc. are frequently used. Taking the example of classifying 22 categories of library collections, this work exploits the application of clustering analysis algorithm in digital library with the use of K-means algorithm and the library collection data like book use rate and collection proportion. The data used in this work comes from the book management system of Jiangnan University. Data processing is done with the SPSS software. 22 categories of library collections are classified into four categories. They are Unnoticed, Shortage in resource allocation, Reasonable resource allocation and Partial-saturation in resource allocation. The high reliability and stability of K-Means clustering algorithm are exploited and verified after applying it to resource analysis of digital library, providing scientific basis for the construction and evaluation of library collection resource.

Introduction

Library, as a center of literature information, generates amount of log data every day, which seems mixed and irregular but valuable in fact [1, 2]. How to regularize these seemingly disordered data and dig out the valuable information so as to serve the digital library construction and also the collection resource construction? Base on this question, taking the example of classifying 22 categories of library collections, this work exploits the application of clustering analysis algorithm in digital library for the purpose of providing support measures for digital library in resource construction, personalized service and reader service, etc., with the use of K-means algorithm and the library collection data like book use rate and collection proportion.

Clustering Analysis

Clustering analysis is a method by which the research objects are classified according to their individual features under certain class definition criterions, and is widely applied to economics, administration and sociology, etc. [3]. By clustering analysis, classification is carried on according to the degree of closeness among individuals. The level of difference is weighed by distance at most of time [4]. Distances between categories, which are defined by various ways, generate different kinds of clustering analysis methods including frequently-used such as hierarchical clustering and K-Means [5].

K-Means clustering algorithm

K-Means is a typical clustering method based on distance. The mean of all data samples in each clustering subset is taken as its cohesion point, with the standard iteration algorithm, which minimizes the square sum of distances between samples and category mean, being employed [6]. The main idea of the algorithm is to divide the data sets into different categories via iteration procedure, in order that the formula function which evaluates clustering performance can be optimized with every created category being compact and independent with each other [7].

Data processing. Divergence of magnitudes in values of each variant should not be tolerated [8]. Normalization treatment includes following methods.

Min-max normalization is to make a linear transformation for the raw data. Its computational formula is as

follow:

$$x_i' = \frac{x_i - \min x_i}{\max x_i - \min x_i}. \quad (1)$$

Z-score normalization normalizes data dependent on the mean value and standard deviation of the raw data. It is applied to the situations in which the maximum and minimum of property A are unknown, or in which outliers(data that exceed value range) exist. Its computational formula is as follow:

$$x_i' = \frac{x_i - \bar{x}_i}{\sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}_i)^2}}. \quad (2)$$

where $\bar{x}_i$ is the mean of sample data.

Algorithm flow. The K-Means algorithm which is applied to the bibliography varieties works as follow:

A sample set $X = \{x_1, x_2, L, x_n\}$, it is divided into k categories according to one's demand, and we obtain the initial clustering centers $c_j(I)$.

$$c_j(I) = \frac{1}{n_j} \sum_{x' \in c_j} x'. \quad (3)$$

Calculate the distance $D(x_i, c_j)$ between each sample x_i and each clustering center c_j , where $i = 1, 2, L, n; j = 1, 2, L, k$. Here the Euclidean distance is adopted, its computational formula is

$$d_{ij} = \sqrt{\sum_{m=1}^n (x_{im} - x_{jm})^2}. \quad (4)$$

where x_{im} is the m th observed value of the i th sample, while x_{jm} the m th observed value of the j th sample, and n denotes variety number of books.

According to the above formula, the distances d_{ij} between any two points are calculated in turn, the larger d_{ij} represents less similarity or more difference between x_i and x_j ; while the smaller d_{ij} shows more similarity between x_i and x_j , and that they have less divergence which makes them belong to one category.

Recalculate k new clustering centers $c_j(I+1)$ according to the new samples partition [9,10]. The K-Means algorithm uses square sum of errors criterion function to evaluate, the computational formula is

$$E = \sum_{j=1}^k \sum_{p \in x^j} \|p - c_j\|^2 \quad (5)$$

Repeat the 2nd and 3rd steps until the cohesion centers $c_j(I)$ become very small or reach the preset restricted condition.

The main idea of K-Means algorithm is to divide the data sets into different categories via iteration procedure, in order that the formula function which evaluates clustering performance can be optimized with every created category being compact and independent with each other. The flow chart of K-Means algorithm is showed in Figure 1.

Figure 1. Flow chart of K-Means clustering algorithm.

Examples of applying clustering analysis algorithm to digital libraries

Clustering analysis is an important method of data mining for finding out data distribution and hidden patterns [7]. Taking the example of clustering algorithm, this work analyzes the use of library collection varieties and the library collection status to find out the internal relations in the data.

Data obtaining. The data used in this work comes from the book management system of Jiangnan University. Data about borrowing situation of 22 categories of books from year 2000 to 2014 is extracted from the system first, and then connect it with the library collection data to make a coherent processing. After processing, the data is showed in Table 1.

Table 1 Library collection using situation

Category names	Category numbers	Collection rates	Using rates	Zero borrowing rates	Updating rates
Marxism-Leninism	A	0.55608	14.05	8.51	2.14
Philosophy	B	3.789945	26.66	6.5	4.21
Social science pandect	C	3.022394	18.16	8.92	3.31
Politics and laws	D	6.849927	15.61	7.6	3.3
...	...	...	...	...	...

Achieve data analysis with K-Means algorithm in SPSS software. The data of collection resource situation is first import into SPSS software, and then choose K-Means Custer, next, giving classifications of books 4 as one demands, the data can be analyzed. This process can be done through Syntax program statements.

Results of clustering analysis. Table 2 shows the variance of categorical variables analysis. The sig value of collection rate, zero lending rate and utilization rate is 0, and statistically significant. The sig value of update rate is 0.097, so it will be eliminated in the following analysis.

The iteration process of clustering is showed in Table 3. It shows that the clustering centers do not change any more after 5 times of iterations, and that the difference values of computational results between the 4th iteration and 5th iteration all become 0, meaning the end of the clustering process.

Table 2 Analysis Of Variance

	Cluster		Error		F	Sig.
	Mean Square	df	Mean Square	df	Mean Square	df
Collection rates	123.206	3	6.454	18	19.090	.000
Using rates	346.919	3	8.147	18	42.585	.000
Zero borrowing rates	21.353	3	1.055	18	20.240	.000
Updating rates	2.346	3	.959	18	2.448	.097

Table 3 Records of iteration process

Change in Cluster Centers				
Iteration	1	2	3	4
1	6.878	7.200	.000	3.941
2	.000	1.073	3.558	.000
3	1.368	1.887	2.402	.000
4	2.174	.770	.000	.000
5	.000	.000	.000	.000

Table 4 represents the clustering result of classifying the library collection varieties into 4 categories, where *Cluster* denotes the category the book variety belongs to, and *Distance* is the distance from each book variety to its belonging category center.

Table 4 Results of clustering

Category name	Cluster	Distance	Category name	Cluster	Distance
Marxism-Leninism	2	3.129	Natural science pandect	1	.587
Philosophy	3	2.169	Mathematical and physical science, chemistry	2	3.178
Social science pandect	2	3.522	Astronomy, Geoscience	1	.876
Politics and laws	2	3.460	Biology science	2	3.109
Military	2	6.314	Medicine, Hygiene	2	3.098
Economics	4	4.153	Agricultural science	1	1.842
Culture, science, education and sports	2	3.479	Industrial technology	4	3.941
Languages, words	3	4.804	Traffic and transportation	2	3.103
Literature	4	3.669	Aeronautics, Astronautics	1	1.457
Art	3	5.670	Environmental science, Security science	2	3.823
History, Geography	2	4.038	Comprehensive books	1	3.600

Result analysis and suggestions for library collection resource allocation. Natural science pandect, Astronomy and Geosciences, etc. are classified into the first category, and can be defined as Unnoticed. Marxism-Leninism, Social-science pandect, etc. are separated into the second category. This category belongs to a reasonable resource allocation. Languages and words, Philosophy, Art are classified into the third category. These book varieties can be defined as Shortage in resource allocation. Purchasing proportion of these varieties of books should be increased. Economy, Literature and Industrial technology are classified as the fourth category. This category can be defined as Partial-saturation in resource allocation. The collection quantity of these varieties of books is at a relative saturation condition and that the relevant purchasing should be decreased.

Summary

Depending on the books using situations, this work analyzes the library collection resource allocation from aspects of books using rate, collection proportion updating rate, etc. with the

application of K-Means clustering algorithm. The high reliability and stability of K-Means clustering algorithm are exploited and verified after applying it to resource analysis of digital library, providing scientific basis for the construction and evaluation of library collection resource. Henceforth, application of other algorithms of clustering analysis to digital library data analysis can be discussed, such as reader classification, personalized service and resource recommendation, providing further decision supports for the development of digital library.

References

- [1] M.L. Lei, X.Y. Lei: Jiangxi Science, Vol. 29 (2011) No.3, p.395. (In Chinese)
- [2] Q. Xia: Information Research, Vol. 194 (2013) No.12, p.115. (In Chinese)
- [3] Information on <http://baike.baidu.com/view/903740.htm?fr=aladdin>. (In Chinese)
- [4] D.H. Liu, N. Li and Y. Huang: Mastering SPSS15.0 Statistical Analysis (Tsinghua University Press, China 2008).
- [5] H.B. Zhu: Applied Statistics and SPSS Application (Electronic Industry Press, China 2011).
- [6] H. Yang: The Application of Clustering Analysis Based on SPSS in Industry Statistical Data (MS., Jilin University, China 2013), p.12.
- [7] H.Y. Ye: Principle and Analysis of Statistics (Wuhan University of Technology press, China 2012).
- [8] R.Y. Li: The Research and Application of Cluster Analysis in Data Mining Algorithm (MS., University of Electronic Science and Technology, China 2012), p.30.
- [9] Y.N. Sun, J.X. Gu: Library Theory and Practice, Vol. 36 (2014) No.3, p.32. (In Chinese)
- [10] C.S. Yang: Journal of Guizhou Normal College, Vol. 28 (2012) No.6, p.11. (In Chinese)