Application of cluster mining and the *apriori* algorithm in the management of library books

Li Ming *

Library, Heze University, Heze 274015, China

Received 15 August 2014, www.cmnt.lv

Abstract

Data mining technology has strong data processing ability. A library is a data resource center in which numerous, potentially correlated data are located. Thus, library management is greatly challenged in terms of determining the internal value of this information and using it effectively. Data mining technology is used for clustering and association rule analyses on library management systems, locating popular books and categories in numerous book resources, and determining the correlations among books. Moreover, the mined data are applied in library management systems to provide new books and recommendation services to readers and to introduce opinions on other analytical methods, thereby enhancing the theoretical research basis for data mining.

Keywords: Data mining technology, clustering mining; association rule mining, library management

1 Introduction

As an emerging inter-discipline, data mining refers to the process of extracting included [1], unknown, and useful knowledge and information from numerous data that are noisy, incomplete, and random. It is the process of locating useful information in a storage database and in other information banks [2]. Data mining effectively processes information and combines traditional mathematical statistics and logical analysis technologies with current artificial intelligence technology to generate a set of comprehensive data analysis methods [3]. One or more mining algorithms are adopted to determine the potential law at work among the data given specific problems and data [4]. In a system, data are mined in a continuous circulation and optimization process. Furthermore, data mining is generally divided into data preparation, the data mining algorithm model, and the evaluation of results [5].

With the rapid development of the economy and of science and technology, electronic books have gradually been incorporated into people's daily lives [6]. However, paper books continue to dominate libraries, which operate under relatively backward management methods and service efficiency [7]. The main system of borrowing and returning is inefficient and cannot meet increasing demands of readers for knowledge [8]. Therefore, library management must adhere to current trends to provide readers with updated and comprehensive functions [9]. Book management systems are an important in university libraries, which are centers of central information processing. Their significance is attributed to the fact that these systems store masses of data and display high information throughput. As a result, information construction in libraries is continuously promoted [8]. The emergence of book management systems is vital to universities in terms of management and decision making in libraries [10]. Book management systems must integrate high efficiency, speed, and convenience. Therefore, the academic field and political circles increasingly emphasize the research into and analysis of library management problems, as well as the determination of an effective management mode. The research on library management must be strengthened to serve readers effectively [11].

Data mining technology is an analytical method of mining correlations among numerous uncertain data to identify hidden valuable information [12]. Its key aim is to detect and classify these correlations and to use clustering or association rules along with other relevant models to process the data. In the process, the potential law can be determined [13]. Data mining technology is usually applied to unforeseeable and invisible information to explain the correlation among raw data. Currently, it is an important method of determining and analyzing numerous academic problems. In the current study, data mining technology is used to investigate library management systems comprehensively based on existing research, which aids in exploring the applications of this technology.

2 Data mining

At present, many data mining analytical methods have been developed. The primary methods include clustering, time-sequence mode, classification, association, and deviation analyses.

* Corresponding author's E-mail: limingli2014@yeah.net

2.1 CLUSTER MINING

Clustering is a common data analysis tool, the purpose of which is to divide numerous collected data points into several categories. The data in each category are highly similar, whereas those under various categories differ considerably. The samples, objects, or variables are classified individually.

Given vectors X_i , X_j , $X_i = \{X_{i1}, X_{i2}, \cdots, X_{im}\}$, and $X_j = \{X_{j1}, X_{j2}, \cdots, X_{jm}\}$ in m-dimension space R_m, the distance between vector X_i and X_j is:

$$\mathrm{dij} = \sqrt{\sum_{k=1}^{m}\left(X_{ik} - X_{jk}\right)}^{2} . \quad (1)$$

The k-means algorithm is typically used in cluster data mining. It is an indirect clustering method based on the measured similarity among the samples. Assuming that *n* objects are clustered and the result is required to generate *k* categories, the algorithm flowchart is as shown in Figure 1 below:

If $K_i = \{K_{i1}, K_{i2}, \ldots, K_{ij}\}$, the calculation formula for assessing the category center is:

$$\mathrm{N} = \frac{1}{\mathrm{n}}\sum_{\mathrm{j}=1}^{\mathrm{n}} \mathrm{K_{ij}}\left(\mathrm{n} \le 1\right). \quad (2)$$

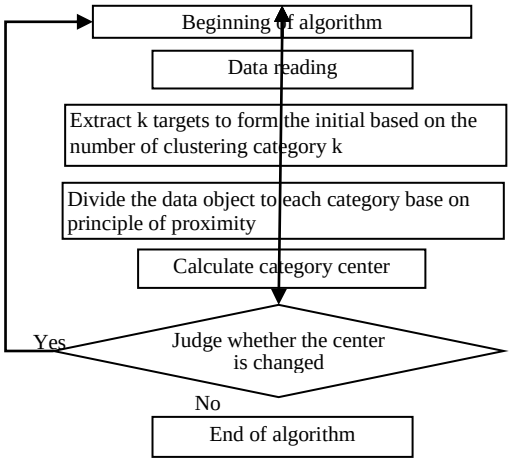

FIGURE 1 Flowchart of the k-means algorithm

2.2 ASSOCIATION RULE MINING

Association rule mining aims to determine the relation or connection among given items by analyzing and recording the set, which reflects the interdependence and correlation of one object with others. If two or more objects are related in a certain way, then one object can be derived through the others. Association rule research is significant in data mining research and is widely applied in various fields.

The data set mined through association rule is called a transaction database, which is recorded as $D = \{t_1, t_2, \ldots, t_k, \ldots, t_n\}$, where $t_k = \{i_1, i_2, \ldots, i_m, \ldots, i_p\}$ and $t_k (k = 1, 2, \ldots, n)$ are known as transactions and $i_m (m = 1, 2, \ldots, p)$ are items. The support degree of the item set $(X \cup Y)$ is considered the support degree of the association rule *XY*, which is expressed as:

$$Support(X \Rightarrow Y) = Support(X \cup Y) . \quad (3)$$

The confidence level of $X \Rightarrow Y$ is written as:

$$Confidence(X \Rightarrow Y) = \frac{\mathrm{Support}\left(\mathrm{X} \cup \mathrm{Y}\right)}{\mathrm{Support}\left(\mathrm{X}\right)} \times 100\%. \quad (4)$$

The *apriori* algorithm is usually applied during data mining under association rule. It is currently the most effective algorithm for this type of data mining. This research also utilizes the layer-by-layer iteration method. The item set with a support degree that exceeds the minimum is called the frequent item set. First, a frequent item set is generated, which is recorded as L1. L1 then produces the item set C1 as the candidate item set to determine the frequent item set 2 L2. Next, the item set C2 generated by L2 is considered the candidate item set to determine the frequent item set 3 L3, and so on. The algorithm is stopped when *k* becomes null.

The algorithm of the core concept is as follows:

(1) L1 = (frequent item set 1)
(2) for(k = −2； $L_{k-1} \neq \Phi$ k++)do begin
(3) Ck = apriori__gen(L_{k-1});// a new candidate item is generated
(4) for all transactions t∈D do begin
(5) Ct = subset(Ck, t); // candidate set contained in t
(6) for all candidates c∈Ct do
(7) c.count^{++};
(8) end;
(9) Lk = {C∈Ct|c.count7>minsup};
(10) end;
(11) return Answer = UkLk ;

3 Application of data mining technology in library management

3.1 FUNCTION OF A LIBRARY MANAGEMENT SYSTEM

A library management system focuses on ordinary users and an administrator. Ordinary users generally refer to readers, for whom a retrieval and recommendation services can be provided for association mining. These services are in addition to book borrowing, returning, shelving, and unshelving services. An administrator can also mine numerous data in the process of administration to conduct clustering and association rule analyses. Cluster analysis mainly emphasizes two aspects: the clustering of readers and of books. According to book

borrowing data, many kinds of books that are popular among the readers in a library can be determined through cluster analysis to increase the number of new books borrowed in categories and to reduce those that have not been borrowed for a long time. This process optimizes the library architecture. During the cluster analysis of book readers, the management system determines the category from which a high number of books have long been borrowed by readers. The system then implements an individualized recommendation service according to user characteristics to improve reading efficiency, which is beneficial for the use of book resources in the library. Association rule mining analyzes and investigates the books and forms of literature in the database of the entire library. It then uses relevant algorithms to determine high correlations among book categories. Readers may opt to borrow simultaneously, and the system generates recommendations of related books when a user searches for a book so he/she can select the necessary books. Figure 2 depicts a structural schematic of the data mining function in a library management system.

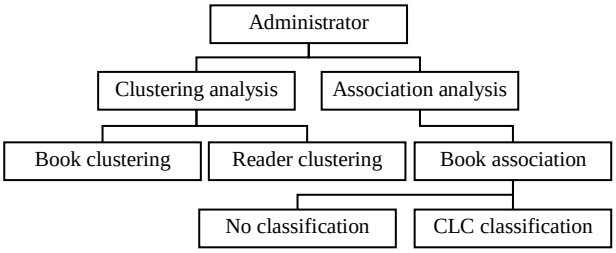

FIGURE 2 Structural diagram of data mining in library management

3.2 APPLICATION OF DATA MINING IN LIBRARY MANAGEMENT

This study mainly employs cluster and association rule mining to mine book and reader data information in the application of data mining to library management systems.

3.2.1 Application of cluster mining to library management

Given a university library as an example, the author considers the student-related book borrowing information in the university (student information, borrowing record, and book details). The borrowing information of some readers is then extracted as a sample by which readers' borrowing behavior can be analyzed. For instance, the number of books borrowed by readers in 2013 is determined, excluding the relevant data and information of readers whose borrowing certificates were canceled in 2013. Finally, 7,148 pieces of borrowing information are obtained. Figure 3 displays the statistical process. The main outputs are detailed reader information and book-borrowing records.

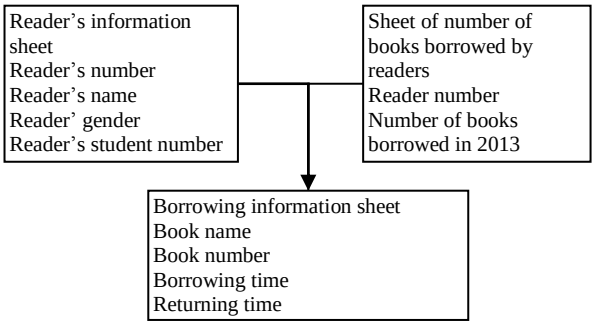

FIGURE 3 Statistics of number of books borrowed by readers

According to the cluster analysis of the statistics on the number of books borrowed by readers, readers with high and low borrowing frequencies are determined systematically, along with their borrowing data. The k-means algorithm is then used in cluster analysis to identify the potential law at work among the data. The information on 20 readers in 2013 is extracted from the library database for calculation, as in Table 1 below:

TABLE 1 Number of books borrowed by readers

Reader number	Number of books borrowed	Reader number	Number of books borrowed	Reader number	Number of books borrowed
201354432	1	201865478	26	206873200	43
201867844	3	201678954	27	201503975	48
204439535	5	201934567	27	203844706	58
208905436	6	201500034	28	204556601	78
201789043	10	204567001	35	204334798	107
201135675	16	207954009	37	204534765	141
201639511	25	205400750	38		

The k-means algorithm is used in the cluster mining analysis of the book borrowing number of 20 readers. Given its limited number, the clustering number k is set as 3, i.e., the clusters is divided into the following three categories: readers with low borrowing frequency, readers with average borrowing frequency, and readers who borrow frequently. Hence, the readers' borrowing table suggests that the clustering tuple is {l, 3, 5, 6, 10, 16, 25, 26, 27, 27, 28, 35, 37, 38, 43, 48, 58, 78, 107, 141}, and $k = -3$, where the first three data, namely, 1, 3, and 5, are selected as the three category centers of cluster mining. $m1 = 1$, $m2 = 3$, and $m3 = 5$.

Euclidean distance is generally used to measure data in cluster analysis. Its expression formula is as follows:

$$d(i,j)=\sqrt{\left|X_{i1}-X_{j1}\right|^2+\left|X_{i2}-X_{j2}\right|+\cdots+\left|X_{im}-X_{jm}\right|^2},$$
$$\text{i},\ j\in\{1,2,\cdots,n\}. \tag{5}$$

$(X_{i1},X_{i2},\cdots,X_{im})\cdot(X_{j1},X_{j2},\cdots,X_{jm})$ corresponds to the target of two m-dimension data.

The k-means algorithm is employed according to the formula above and to the central distance, and the results obtained are presented in Table 2.

TABLE 2 Result of clustering mining

Number of iteration	m1	m2	m3	k1	k2	k3
1	3	27.6	108.6	{1,3,5, 6,10}	{25,26,27,27,2 8,35,37,38,43, 48,58}	{78,107,141}
2	5	34	108.6	{1,3,5, 6,10, 16}	{25,26,27,27,2 8,35,37,38,43, 48,58}	{78,107,141}
3	6.8	35.6	108.6	{1,3,5, 6, 10,16}	{25,26,27,27,2 8,35,37,38,43, 48,58}	{78,107,141}

The clustering result does not change under multiple iterations. In this case, the clustering is ended.

An actual cluster mining experiment is conducted on the readers' borrowing data to demonstrate the reliability of the clustering results. The readers are classified into the following three categories: lazy, ordinary, and active readers. A statistical cluster analysis is then conducted on them to obtain the results exhibited in Table 3.

TABLE 3 Statistical results of various readers

Clustering classification	Number of recording	Proportion in total number of recording	Average of number of borrowing	Total number of borrowing	Library utilization
Lazy readers	2230	31.20%	6.9	15387	7.05%
Ordinary readers	4010	56.10%	31.1	124711	57.11%
Active readers	908	12.70%	86.2	78269	35.84%
Total	7148			218367	

The k-means algorithm results in Table 2 suggest that *k1*, *k2*, and *k3* represent three tuples, respectively, each of which represents the similar borrowing data. Tuple *k1* {1,3,5,6,10,16} indicates that the readers in that group seldom borrow books; therefore, they are lazy readers; *m1* denotes an average of 6.8, which accounts for 30% of the total readers. These readers seldom utilize the library for reading or borrowing. The readers in *k2* are active and frequently borrow books from the library, with an average borrowing number of 108.6. Table 3 suggests that the calculation results of the k-means algorithm are consistent with the actual mining analysis results, in which inactive readers account for a large proportion. To improve the book utilization in a library and to cultivate students' reading habits in actual management applications, the reason why students do not like reading can be investigated on the basis of cluster mining. Thus, corresponding countermeasures can be implemented.

3.2.2 Application of association mining to library management

The Apriori algorithm is applied to library data for association rule analysis. The one-time borrowing record data of five readers were randomly extracted on December 4, 2013.

TABLE 4 One-time borrowing information of 5 readers on 04/12/2013

Borrowing number	Reader number	Category of books borrowed
K1	223423311	G1,B0,TP,I2,B83,G4,H0,TU
K2	201239002	H3,I1,I3,Q56,TP,X4,TU
K3	204377033	C87,H3,I3,J60,K5
K4	201189070	H3，I1，K80，TP
K5	201599077	C80，H0,I34,TB

The above data are preprocessed for convenient association analysis. The set is divided into large categories; for example, reader *K1* is denoted by K1{223423311} = {B,G,H,I,T} as presented in Table 5.

TABLE 5 Preprocessed borrowing information of five readers

Borrowing number	Reader number	Category of books borrowed
K1	223423311	B,G,H,T,T
K2	201239002	H,I,Q,T,X
K3	204377033	C,H,I,J,K
K4	201189070	H,I,K,T
K5	201599077	C,H,I,T

The *apriori* algorithm is used to process the data of five readers and determine the correlation.

The binary matrix of the items in set 1 is constructed after primary scanning, where 0 represents the existence of a correlation among the data and 1 denotes no correlation.

$$P=\begin{array}{c} \\ K1 \\ K2 \\ K3 \\ K4 \\ K51 \end{array}\begin{array}{c} \begin{array}{cccccccccc} B & C & G & H & I & J & K & Q & T & X \end{array} \\ \begin{bmatrix} 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \end{array}.$$

The matrix line is interchanged with a column to obtain matrix P1.

$$\begin{array}{c} \begin{array}{ccccc} K1 & K2 & K3 & K4 & K5 \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix} \end{array} \quad P1= \begin{array}{c} \begin{array}{ccccc} K1 & K2 & K3 & K4 & K5 \end{array} \\ \begin{bmatrix} B & 1 & 0 & 0 & 0 & 0 \\ C & 0 & 0 & 1 & 0 & 1 \\ G & 1 & 0 & 0 & 0 & 0 \\ H & 1 & 1 & 1 & 1 & 1 \\ I & 1 & 1 & 1 & 1 & 1 \\ J & 0 & 0 & 1 & 0 & 0 \\ K & 0 & 0 & 1 & 1 & 0 \\ Q & 0 & 1 & 0 & 0 & 0 \\ T & 1 & 1 & 0 & 1 & 1 \\ X & 0 & 1 & 0 & 0 & 0 \end{bmatrix} \end{array}$$

Five readers are present according to the minimum support degree (*Minsup* = 60%) of the recognized association rule. Thus, $b = 5 \times 60\% = 3$. Each line of the matrix is then calculated to determine the number with a correlation of 1. If the number of 1 is less than *b*, this item does not belong to the frequent set and is deleted, otherwise, the number of 1 is reserved.

The *P1* matrix is eliminated to obtain the items that are finally reserved and to establish the new model matrix *P2*, i.e.,

$$P2=\begin{array}{c} \\ H \\ I \\ T \end{array}\begin{array}{c} \begin{array}{ccccc} K1 & K2 & K3 & K4 & K5 \end{array} \\ \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 \end{bmatrix} \end{array}.$$

The following is the elimination and calculation of the logic of every two lines for matrix *P2*:

$$P3 = \begin{matrix} \\ HI \\ HT \\ IT \end{matrix} \begin{matrix} K1 & K2 & K3 & K4 & K5 \\ \end{matrix} \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 \end{bmatrix}.$$

Ten *P3*s are eliminated to derive the final frequent set:

$$P4 = \begin{matrix} K1 & K2 & K3 & K4 & K5 \\ HIT & & & & \end{matrix} \begin{bmatrix} 1 & 1 & 0 & 1 & 1 \end{bmatrix}.$$

The algorithm is then concluded, and the final frequent set is {H,I,T}.

H, *I*, and *T* operate under very strong correlation rules. Therefore, books *H*, *I*, and *T* are the books most likely borrowed by readers. When the book borrowing category is H, 60% of the readers may borrow from categories *I* or *H*, and vice versa. In practice, *H* represents the language books, *I* denotes literature and social science books, and *T* indicates industrial and technical books, which are closely related to the learning needs of the readers. Therefore, the chances that readers borrow such books are high.

4 Conclusion

The use of algorithms for processing and analysis has significantly improved the efficiency of data mining with the wide applications of computers, and the application of data mining has gradually been widened in various fields. Data mining technology is used in the clustering and association rule analyses of library management systems. It is also applied to determine the popular books and categories in numerous book resources and the correlations among books. The mining results are incorporated into library management systems to generate new book categories and to provide recommendation services to readers. Data mining technology is constructed through computer programming and data algorithms; the influences of subjective factors are overcome during analysis. The obtained results are objective and reliable.

References

[1] Srechko Natek, Moti Zwilling 2014 Student data mining solution–knowledge management system related to higher education institutions. *Expert Systems with Applications* **41**(14) 6400-07

[2] Neda Abdelhamid, Aladdin Ayesh, Fadi Thabtah 2014 Phishing detection based Associative Classification data mining. *Expert Systems with Applications* **41**(13) 5948-59

[3] *Deleted by CMNT Editor*

[4] Kyunglag Kwon, Daehyun Kang, Yeochang Yoon, Jong-Soo Sohn, In-Jeong Chung 2014 A real time process management system using RFID data mining *Computers in Industry* **65**(4) 721-32

[5] *Deleted by CMNT Editor*

[6] McVay M R, Kokoska E R, Jackson R J, Smith S D 2008 Throwing out the "grade" book: management of isolated spleen and liver injury based on hemodynamic status *Journal of Pediatric Surgery*, **43**(6) 1072-6

[7] Aggarwal R, Simkins B J 2001 Open book management optimizing human capital *Business Horizons* **44**(5) 5-13

[8] Davis T R V 1997 Open-book management: Its promise and pitfalls *Organizational Dynamics* **25**(3) 7-20

[9] Barbalho H, Rosseti I, Martins S L, Plastino A 2013 A hybrid data mining GRASP with path-relinking *Computers & Operations Research* **40**(12) 3159-73

[10] Hong Tzung-Pei, Lee Yeong-Chyi, Wu Min-Thai 2014 An effective parallel approach for genetic-fuzzy data mining *Expert Systems with Applications* **41**(2) 655-62

[11] Shyur Huan-Jyh, Jou Chichang, Chang Keng 2013 A data mining approach to discovering reliable sequential patterns *Journal of Systems and Software* **86**(8) 2196-203

[12] Strohmeier Stefan, Piazza Franca 2013 Domain driven data mining in human resource management: A review of current research. *Expert Systems with Applications* **40**(7) 2410-20

[13] *Deleted by CMNT Editor*

Author

Li Ming , 1974.01,Shandong Province, PR China

Current position, grades: the Intermediate controller of Library, Heze University, PR China.
University studies: She received her Bachelor's degree from Liaoning University in PR China.
Scientific interest: Her research interest fields include Library Management 、 Library and Information Science.
Publications: more than 5 papers published in various journals.
Experience: She has teaching experience of 20 years, has completed 2 scientific research projects.