A Study on Associations between Different Classifications of Library Circulation Data

Wen Ru and Zhanhong Xin
School of Economics and Management, Beijing University of Posts and Telecommunications, Beijing, China
Email: ru_wen@163.com, xinzhanhong@263.net

Di Gan and Jun Xing
Information and Technology Department, National Library of China, Beijing, China
Email: {gandi, xingjun}@nlc.gov.cn

***Abstract*—On the base of analyzing data mining techniques based on association rules, a method of applying Apriori association rules to classify and analyze library circulation data is discussed. In order to remedy the defections of Apriori association method, optimizing algorithm is presented and explained in detail. By using the historical circulation data gathered from the Library Integration System of NLC (National Library of China), this paper studies the circulation data collecting, processing and classifying methods and analyzes the actual circulation data of NLC from 2012 to 2014 with the java programs. Finally, some suggestions for improving the service quality and meeting the readers' reading requirements are given through the analysis results.**

***Index Terms*—classifications of circulation data, association rules, optimizing Apriori algorithm, empirical study**

I. Introduction

For large libraries with a vast amount of inventory, how to organize the collections with a proper classification method and shelve them in an order that caters better to the readers' reading hobbies and habits is a subject worthy to study. What is seen more usually in libraries is to shelve the collections by conventional classification methods, which may appear easy for a reader to retrieve what he or she needs for materials, but actually may not always be the case. More often than not, what a reader wants to borrow are not limited to only one book or one category of books, but various books ranging over much wider areas.

For example, a computer fan may also want to borrow materials on mathematics when seeking computer related materials. So, when by conventional classification method, the reader may have to search two separate reading rooms for the access to the materials he or she needs. Suppose that we can foresee the correlation between the different categories of materials each time our readers need, and shelve them accordingly, library service would be more humanized and more efficient for users to target their materials [1]. In this paper, we apply the Association Rules algorithm in data mining to the analysis of library circulation data so as to obtain the correlation rules from readers' borrowing and reading behaviors of different material categories, and thus to acquire a new optimized shelving scheme that could maximize the convenience for readers with the Apriori algorithm as much as possible.

II. Research of Data Mining and Association Rules Algorithms

A. Association Rules Algorithms in Data Mining

To undertake the research on readers' borrowing and reading behaviors, we must begin with data mining based on the mass data of such behaviors. Data mining is the process of extracting potentially useful information and knowledge from a huge amount of data, which, though fragmentary, noisy, fuzzy and randomized, always harbor unpredictable value [2].

Methods in terms of data mining include those of statistics, machine learning, neural network, database and so on, among which the Association Rules is an important one adopted prevailingly. The association rules method, put forward by R. Agrawal et al in 1993 [3], was initiated to solve problems such as transactional database analysis, etc. It is a technical method intended to identify the universal rules borne in objects through discovering potential associations between items in databases and finding unknown reliance relations laid within mass data.

Of all the various algorithms that have been developed based on Association Rules, AIS algorithm [4], DHP (Direct Hashing and Prune) algorithm [5], FP-growth algorithm [6], [7] and over weighted association rules are the most typical[8].

AIS algorithm is the first proposed association-rule-based algorithm, and its realization process is to generate data sets when scanning database and in the mean time take the count of them [4]. On reading one record, the algorithm searches within the record for the frequent data itemset generated after the previous search, and if found, all such frequent data itemsets and other data items within this record will be formed into a new candidate data itemset by an expanding operation.

Manuscript received March 23, 2015; revised April 20, 2015.

doi: 10.12720/jait.6.3.173-177

And Apriori algorithm utilizes priori knowledge in the process of generating candidate data itemsets, which was operated separately from the calculation of data sets. Therefore, it is provided with two advantages as follows: (1) the generated candidate data itemsets are relatively more targeted; (2) Candidate data itemsets no longer need to be generated repeatedly according to records in the database, instead, they are one-off generated during the process of searching for frequent data itemsets k(which indicates K times of looping) by the frequent data itemsets k-1, which was generated from the previous loop.

While DHP algorithm uses Hash technique, and therefore could more effectively generate candidate frequent data itemsets [5], which in return could remarkably reduce the size of database to be scanned and improve searching speed in the later period of the running of the algorithm by virtue of their features.

Apriori algorithm, which was put forward by Agrawal and Imielinski [3], is the most classical one of all the algorithms mentioned above. By using support and confidence, Apriori algorithm discovers strong association rules between transactions, and thereby became popular in large databases of retail trading, website development, medicine, financial investment, library management system and so on. Apriori is succinct and easy to implement, that's why we are using this algorithm to analyze the data mining of readers' borrowing and reading behaviors in the following case.

To use Apriori algorithm, we should first learn its relevant definitions and how they are implemented, the following being its brief introduction:

B. *Definition of Apriori Algorithm*

Definition1: Dataset D set as the general transactions database, $D=\{t_1,t_2,\cdots,t_n\}$, t as transactions, assigning each transaction with a unique identification, TID, one transaction have many items, $t_k=\{i_1,i_2,\cdots,i_m\}$, i as item.

Definition 2: Suppose $I=\{i_1,i_2,\cdots,i_m\}$ is a set of all items in D, any subset X of I is called Itemset of D. In $|X|=k$ we call X the k-Itemset.

Definition 3: If *X,Y* are itemsets, and $X\cap Y=\Phi$,then implication $X\Rightarrow Y$ is called an association rule, X,Y are called the premise and conclusion. The support level of itemset $X\cup Y$ is called support level of this association rule, denoted by $\text{support}(X\Rightarrow Y)$. There two type of support level, one is called absolute support while another relative support. Absolute support is the count of transactions included by itemset $X\cup Y$ in database D while relative support is the ratio of the count of transactions and the complete transactions in database D. that is

$$\text{support}(X\Rightarrow Y)=P(X\cup Y)=\frac{|\{T:X\cup Y\subseteq T,T\in D\}|}{|D|}$$

When an itemset's support level is no less than the minimum support level (minsupport) designated by the user, we call the itemset the frequent itemset. The confidence level of association rule $X\Rightarrow Y$ is denoted by:

$$\text{confidence}(X\Rightarrow Y)=P\left(\frac{Y}{X}\right)=\frac{\text{support}(X\Rightarrow Y)}{\text{support}(X)}\times 100\%$$

The minimum confidence level designated by user is denoted by: minconfidence.

Definition 4: If $\text{support}(X\Rightarrow Y)\geq \text{minsupport}$, and $\text{confidence}(X\Rightarrow Y)\geq \text{minconfidence}$, call association rule $X\Rightarrow Y$ is a strong rule, otherwise it's a weak rule.

Support level and confidence level are two major concepts in describing association rules. The former measures the statistical importance of association rules in the whole dataset, and the latter measures the credibility of association rules. In general, only those association rules with both high support level as well as high confidence level are of interest and help for users.

III. Building and Optimizing an Association-Rule Module Based on Apriori Algorithm

Some defections, described below, are found during the actual application of Apriori algorithm:

1)Transaction database is required to be repeatedly scanned;

2)It is inapplicable to the data mining of dense datasets;

3)The association rules probably generated could be excessively huge.

Therefore, we need to have the algorithm optimized before applying it to the mining of library borrowing and reading data.

1)For transaction database D with a known scale, the support level of any item I is irrelevant to transactions which have smaller scale than the amount of $|I|$, and can thus be deleted when the database scanning times equal the amount of $|I|$.

2)In candidate itemset k, if an itemset does not contain any k-1 itemset, then it cannot be a frequent itemset, so we can delete records of such transactions during the k times scanning, hence the reduction of the records amount needed to be scanned the next time.

The two rules above can be summarized as "a subset of a frequent itemset must be a frequent itemset" and "the superset of non-frequent itemsets must be non-frequent". Please refer to Fig. 1 as the flow chart of optimization.

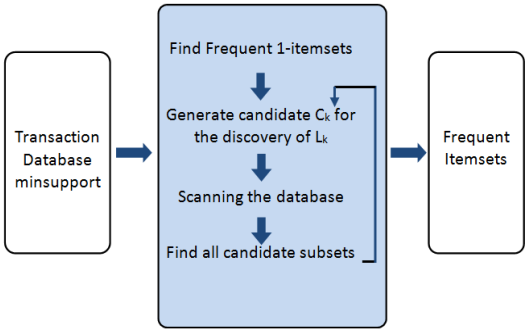

Figure 1. Optimized Apriori association rules procedures

Through template frequent itemsets, first, add the first item of (k-1)-Itemset into the template frequent itemsets. Then, add the other itemset with different last item into the template frequent itemsets, create k-Itemset and calculate the support. If the support larger than minsupport, then create the frequent itemset and save, otherwise delete the itemset. According to the order cycle, the iteration continues until all frequent itemsets are generated.

IV. Implementation of Apriori Algorithm

The implementation of Apriori algorithm is shown in Alogrithm 1 and Alogrithm 2:

Input: Transaction database D; Minimum support level minsupport.

Output: Frequent dataset L in D.

a. the main function of apriori is seen in Algorithm1.

Algorithm 1. Apriori main functon

```
1)L1=all frequent 1-datasets;
2)For ( k= 2; Lk-1≠φ;k++ )
3)    Ck= Apriori_gen(Lk-1,minsuppor t) ;
4)    For all T∈D do
5)        Ct=Subset(Ck ,T);
6)        For all c∈Ct do
7)            c. count++ ;
8)        End For
9)    End For
10)   Lk= {c∈Ck| support (c) >= minsupport}
11)End For
12)Return L={all Lk}
```

Step 1 of the Apriori is to find frequent itemset L_1. In steps 2-11, $L_{k-1}$is used to generate candidate C_k in order to find L_k. In step 3, the Apriori_gen procedure generates candidate and eliminates candidates with non-frequent subsets. Step 4 scans the complete database. Step 5 uses Subset function to find all candidate subsets in the transaction. Steps 6 to 8 accumulate the count for each of such candidates. At last, all candidates that meet the minsupport will form the frequent dataset L.

b. the algorithm of procedure of Apriori_gen

Procedure Aprior_gen generates candidate itemset C_k during the kth times iteration by L_{k-1}(frequent k-1 itemsets), and the algorithm of Aprior_gen is described in Algorithm2.

Algorithm2. procedure of Apriori_gen

```
1)For each itemset l1∈Lk-1
2)   For each itemset l2∈Lk-1
3)       if(l1[1]=l2[1])∧(l1[2]=l2[2])∧…∧(l1[k-2]=
4)          l2[k-2])∧(l1[k-1] < l2[k-1]) then{
5)             c=l1[1],l1[2],…, l1[k-1], l2[k-1]
6)         };
7)       Ck= Ck∪c;
8)       For (each subset s which contains k-1 items)
9)          if (s does not belong to Ck-1)
10)             delete c from Ck;
11)      End For
12)  End For
13)End For
14)Return (Ck) ;
```

The key to increasing Apriori's efficiency is to generate a relatively small candidate itemset C_k, in other words, to try to avoid generating and calculating itemsets which have no possibility of becoming frequent itemsets. By this way, the computation complexity is much simplified, thus greatly improving association-rule-based mining efficiency.

V. Emperical Study on the Mining of Classifi-cations of Circulation Data

The circulation data of a library are derived mainly from library integrated systems. Records are generated in these systems when a reader borrows or returns a book, and these records usually contain information of the readers, books or other materials borrowed, operators and time. Here we take the National Library of China as an example, and have extracted its historical data in two years as basic data for analysis from NLC's Integrated Library System: ALEPH500. Since extraction of historical circulation records without selection will cause unnecessary redundancy, we first made an extraction of data which are in accordance with some linkage based on readers' borrowing and returning and proceed with the selection from within the result. For example, only those readers with loan records of 3 or more copies of materials are taken into account and their personal historical information and circulation data extracted.

A. Obtaining Circulation Data

NLC's ALEPH500 system contains all information concerning current and historical loan. Current loan information consists of item ID, reader's ID, loan date and time, due date and time, circulation staff, etc. While historical loan information consists of item identification, reader's ID, loan date and time, due date and time, return date and time, operating staff, etc. These data are mainly stored in two tables: Z30 and Z36H.We extracted 100,000 records in total from the data between March 10, 2012 and March 9, 2014 randomly for our analysis and study.

B. Obtaining Readers' Information

At present, readers usually borrow or return books through RFID devices. As a prerequisite for this, each reader needs to register a reader's card and pay corresponding deposit. The information of whoever has registered an NLC reader card will be written into the ALEPH system as soon as the card is registered. The records of a reader's information consist of the reader's name, birth date, gender, reader type, status, contact, address, etc. The readers' information records are stored mainly in four database tables: Z303, and Z308 linked by reader's ID. The information of books such as title, publication, call number, barcode are stored in table Z30, while the historical circulation data are stored in Z36h.

The following SQL statements can be used to get the classification numbers of books loaned in certain time frame:

Select *substr(t3.z308_key,3,18),t2.birth_date, t1z30_barcode, t1.z30_call_no,t.z36h_loan_date* ***from*** *z36h t,z30 t1,z303 t2,z308 t3* ***where*** *t.z36h_loan_date <'20140309'* ***and*** *t.z36h_loan_date>'20120310'* ***and*** *t.z36h_key=t1.z30_key* ***and*** *t.z36h_id=t2.z303_key* ***and*** *t.z36h_id=t3.z308_id* ***and*** *substr(t3.z308_key,0,2)= '01'* ***and*** *t1.z30_sub_library='ZWWJ';*

In above statements, z36h_loan_date, representing loaning date and time, is 8 bit numeric type with the format of YYYYMMDD, while z30_call_no is the classification number of the books loaned, and z30_barcode the barcode of the books loaned. The specific information needed can be modified depending on actual circumstances. For the detailed implication of the fields, the User Manual of ALEPH system can be consulted as reference.

C. Program Coding and Data Processing

We have developed a relevant mining and processing program with java language, based on the optimized Apriori algorithm.

1) Treatment of the classification numbers

NLC at present uses CLC (Chinese Library Classification), which has 22 major classifications in total, to classify its collection[9]. Considering the amount of books under TP is remarkably huge, we here will specifically treat the second tier classification TP under class T as a major class. Thus the processing of classification numbers of books will be performed under 23 categories [10].

2) Format of data storage

The classified circulation data that are obtained need to be further processed for the analysis of the program, which means we need to join all the classification numbers of the books that a reader has loaned, and separate these numbers with a semicolon. As a result, a piece of input data consisting of two fields(reader's ID and classification number string) is formed.

D. Analysis of Data Mining Results

We extracted 100,000 circulation data records randomly in resent two years for analysis, and experimented with varied support level and confidence level thresholds to perform the data mining separately based on association rules. The results are shown in Table I:

The mining result is a list of all non-void subsets whose confidence level are greater than the minimum confidence level. When the minimum support levels are the same, so are their frequent itemsets such as the absolute misupport is 200, 500 or 1000. When the minimum confidence level increases from 0.3 to 0.9, the number of association rules decreases correspondingly, and as the same, when the minimum support level increases from 200, 500 to1000, the association rules decreases too. Once the minimum support level is not changed, the content of frequent sets are not changed even with different minimum confidence. If the minimum confidence level is high, it may not get any association rules in result. When the minconfidence is 0.9, three level of minsupport all does not get any association rules.

TABLE I. THE RESULT TABLE OF DATA MINING

absolute minsupport	min-confidence	Frequent sets(partial)	Association rules
200	0.3	I;: 7206 F;: 4856 T;: 3919 TP;: 3754 K;: 3740 H;: 3529 R;: 3214 B;: 2824 D;: 2761 J;: 2187 G;: 2062 C;: 1271 O;: 716 P;TP; 525 I;K; 415 … K;R;: 200	B;K;->I; 0.69091 B;F;->I; 0.61756 F;I;->B; 0.58445 B;I;->K; 0.58312 B;I;->F; 0.55754 I;K;->B; 0.54939 ….
200	0.6	same as above	B;K;->I; 0.69091 B;F;->I; 0.61756
200	0.9	same as above	N/A
500	0.3	I;: 7206 F;: 4856 T;: 3919 TP;: 3754 K;: 3740 H;: 3529 R;: 3214 B;: 2824 D;: 2761 J;: 2187 G;: 2062 C;: 1271 … …	N/A
500	0.6	same as above	N/A
500	0.9	same as above	N/A
1000	0.3	I;: 7206 F;: 4856 T;: 3919 TP;: 3754 K;: 3740 H;: 3529 R;: 3214 B;: 2824 D;: 2761 J;: 2187 G;: 2062 C;: 1271	N/A
1000	0.6	same as above	N/A
1000	0.9	same as above	N/A

Some conclusions can be draw by analyzing the above results, classification I (Literature) has the highest support level as 7206, followed successively by class F (Economy) ,T (Technology), TP (Automation & computer technology) and K (History & geography). Considering TP is a classification separated from T, their sum as 77673 is actually higher than classification I as 7206. It can thus be seen that, in the National Library of China, technology, literature, economy and history are the most loaned categories, mainly because as a comprehensive library, the demands of NLC readers on these subjects are the most.

Other conclusion can also seen that most readers who have borrowed materials of classification B (philosophy, religion) and K (history & geography) exhibit interesting

in materials of classification I as well because the confidence of $B, K \Rightarrow I$ is the highest value of 0.69091. As second high confidence of 0.61756 is $B, F \Rightarrow I$, there are other types of association rule with the confidence higher than minconfidence including $F, I \Rightarrow B$, $B, I \Rightarrow K$, $B, I \Rightarrow F$, $I, K \Rightarrow B$ etc. So according the result, it can be drown that B, K, F and I are the most intensely linked classifications.

VI. CONCLUSION

This paper carries out an analysis on the classified information of the historical circulation data with the optimized Apriori algorithm. It also analyzes the association levels of different classifications by studying the internal correlations of huge amounts of historical circulation data. The data mining process in the case study of NLC showed that as a comprehensive library, a higher proportion of NLC's readers exhibited inclination of borrowing materials on literature, history and philosophy, and association levels between the three classifications are higher too. So, on the one hand, we should take into consideration the purchase of more materials on such subjects; on the other hand, materials should be organized in a way easier for the readers to fetch, so as to further the convenience of reading for users.

Classification information extracted from circulation data reflects only one side of readers' behaviors. How to take into account more perspectives for further study and analysis on library readers' behaviors and preferences such as the information about readers' identity, age, gender and loan time, is in need of continuous focus. Also, research on applying some new statistical methods and mathematical theories, such as the small-world theory to the mining of massive library circulation and readers' data is a direction calling for further exploration.

REFERENCES

[1] J. J. Xie and A. P. Ding, "The application of association rules in library management," *Journal of Henan University (Natural Sciences Edition)*, vol. 38, no. 4, pp. 422-424, 2008.

[2] J. W. Han, M. Kamber, M. Fan, X. F. Meng, et al. *Concepts and Technologies of Data Mining*, China Machine Press, 2001, pp. 213-234.

[3] R. Agrawal and T. Imielinski, and A. Swami, "Mining associations between sets of items in massive databases," in *Proc. the ACM SIGMOD Int'l Conference on Management of Data*, Washington D.C., May 1993, pp. 207-216.

[4] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," in *Proc. 20th VLDB Conference on Very Large Data Bases*, Santiago, Chile, vol. 23, no. 3, 1994, pp. 21-30.

[5] J. D. Holt and S. M. Chung, "Parallel mining of association rules from text databases," *Journal of Supercomputing*, vol. 39, no. 3, pp. 273-299, 2007.

[6] J. W. Han, J. Pei, Y. W. Yin, and R. Mao, "Mining frequent patterns without candidate generation: A frequent-pattern tree approach," *Data Mining & Knowledge Discovery*, vol. 8, no. 1, pp. 53-87, 2004.

[7] S. K. Tanbeer, C. F. Ahmed, B. S. Jeong, et al., "Efficient single-pass frequent pattern mining using a prefix-tree," *Information Sciences*, vol. 179, no. 5, pp. 559-583, 2009.

[8] K. Zhang, X. G. Zhang, et al., *Data Mining Algorithms and Engineering Application*, China Machine Press, 2006, pp. 78-85.

[9] W. Zhang, "An empirical analysis of the relevance knowledge discovery of the reader's lending behavior," *Work and Research of Library*, vol. 12, pp. 38-41, 2010.

[10] Q. F. Liu. "The empirical analysis of the association rules of readers of university lending behavior: Taking Wuhan university of science and technology as an example," *Academic Library and Information Tribune*, vol. 2, pp. 16-18, 2013.

Wen Ru is currently a Ph.D. candiadate of School of Economics and Management of Beijing University of Posts and Telecommunications. He works at National Library of China as a senior engineer and his research area includes information management and digital library. He has undertaken a research project of Ministry of Culture of China and three research projects of NLC.

Zhanhong Xin is currently a professor and Ph.D. supervisor of School of Economics and Management at Beijing University of Posts and Telecom- munications. His key area of researching is Applications of OR and he has published dozens of papers in various periodicals and proceedings.

Di Gan is currently a senior engineer at the Information Technology Depart- ment of National Library of China. His research interests include network security and artificial intelligence. He played a major role in leading the establishment of several sub- projects in the National Digital Library Project.

Jun Xing is currently deputy director of the Information and Technology Department at National Library of China. She has undertaken a number of research projects and huge information systems building projects. Her research area includes computer technology, digital library and cloud computing.