Application of Improved Association-rules Mining Algorithm in the Circulation of University Library

Tingting Xia[a, *], Yingjun Liu[b]

Ningbo University of Finance &Economics, Ningbo City,Zhejiang Province, China

acircle1213@126.com, bliuyj@nbdhyu.edu.cn

*corresponding author

Keywords: Library, data mining, association-rules

Abstract: The rapid development of network technology has brought great influence to library circles, How to reposition and retain readers of Libraries in the new period has aroused the thinking of industry. The consensus view is that libraries need to tap the needs of users to promote reader reflux. This paper is based on data mining technology and association-rules mining technology, by using the improved Apriori algorithm to mine frequent item sets in transaction databases for circulating information in library automation system. Through data mining and analysis, we can grasp the interests and needs of readers, offer some meaningful and practical results for managers, and provide decision-making suggestions on book purchasing, subject construction and collection distribution. Based on data mining technology and association rule mining technology, taking circulation information in library automation system as a sample, this paper uses improved Apriori algorithm to mine frequent item sets in transaction database. Through data mining analysis, it can grasp interests and needs of readers, and provide decision-making suggestions for library purchase, subject construction and collection distribution.

1. Introduction

Library automation system has a history of several decades, from traditional manual processing to computer information processing. This transformation has brought huge information. Data mining technology is a new digital technology. Practice has proved that data mining technology has a broad application background in the digital era. Lone [1] proposed that libraries can use data mining technology to extract strategic information and solve some critical problems. At the same time, the library realizes intelligent data mining. Taking Syracuse University Library as an example, Dr Nemati[2] introduced how to use data mining technology to help library make management decisions. After cleaning and screening the internal and external information of the library system, the data warehouse is established, and the results of analysis are visualized by using multi-class data mining algorithms, which can provide decision-making reference for Library and school administrators. Guo[3] and Cui[4] analysed the Apriori algorithm in association-rules mining, and proposed a new algorithm called NApriori algorithm.

NApriori algorithm uses frequent item sets to reorganize transaction databases and reduces the number of database scans to improve efficiency. Deng[5] applied the improved algorithm to bibliographic recommendation service.

Data mining technology is the key technology in the field of computer and artificial intelligence. Applying this technology to library management, by mining useful information, providing readers with perfect service and scientific basis for decision-making has become the key research direction of Library management.

2. Data Mining Technology

2.1. Flow of improved Apriori algorithm

Since the development of data mining technology, its technical categories have been very large.

  DOI: 10.25236/icbdai.2018.010

Generally speaking, according to its functional attributes, we can divide them into two families: descriptive analysis and predictive analysis. Descriptive analysis includes association analysis, sequence analysis and clustering analysis. The family of predictive analysis consists of two categories: Classification and Statistical Regression[6].

2.2. Apriori algorithm

The core idea of Apriori algorithm is to find frequent item sets by reading candidate item sets in the database N times. Simply speaking, it generates frequent item sets by means of iteration. There are two steps of each iteration: generating candidate sets; computing and selecting candidate sets.

In the first iteration, the database is read first, and a single candidate item set is used as each frequent item sets. The list of candidate item sets is gradually produced. Redundant or duplicate candidate item sets are deleted. Five candidate items sets 1, namely {a}, {b}, {c}, {d}, {e}, are obtained. Their support degree is determined according to the frequency of candidate item sets appearing in the transaction database, and frequent items whose value is less than the minimum support degree are deleted. Set.

3. Improved Apriori algorithm

3.1. Flow of improved Apriori algorithm

Step1: Set appropriate parameter values, scan the whole transaction database, record and count each transaction code, delete redundant or redundant items in the scanning process, and get frequent item sets.

Step2: The candidate item sets are generated by intersection or union of frequent item sets, and their transaction sets are computed to eliminate redundancy and generate the maximum frequent item sets.

Step3: The maximum frequent item sets are generated by recursion layer by layer, except for frequent item sets.

Step4: Loop until frequent item sets cannot be generated.

Figure 1 describes the process of the improved Apriori algorithm in code form:

```
Input: transaction database D; Minimum support threshold min_sup
Output: frequent item sets L of D
L1=find_L1(D, min_sup);
    for(k=2; Lk-1≠∅; k++){
    L'k-1= Lk-1;
//Calculate the number of occurrences of each item in Lk-1 and record the frequent sets
    For each item x∈ Lk-1
      For each field f∈ x
      { f. count ++; If[f. count]=x;}
    //Delete items in L'k-1 that cannot produce frequent sets K
    Delete(L'k-1, If[f. count < k-1);
    For each item x∈ L'k-1
      For each field y∈ L'k-1
    //Find frequent items that can be connected
    if(x1=y1∧x2=y2∧...∧xk-2=yk-2∧xk-1=yk-1)
    {n=0;
    //Pre-ordering facilitates quick identification of the number of identical transactions
    For (i=1,j=1,i≤|Tx|, j≤|Ty|;)
        if(Tx[i] < Ty[j])
            i++;
        else if(Tx[i] > Ty[j])
            j++;
            else
            { n++;
              Tx-y= Tx-y∪ {Tx[i]}
              i++;
              j++;
            }
        If(n≥N×min_sup)      //Find the next frequent item sets
        { Tx-y=Tx-y;
          Lk = Lk∪ {x∞y}
        }
        retun L=∪ Lk;
        //Generating 1-item frequent item sets
    Procedure find_ L1 (D, min_ sup);
    L1=get_ item (D)          //items to remove transaction databases
    For each transaction t∈ D
    { N++;
        For each item x∈ t
        { x. count++;
          Tx [x. count] = td;
        }
    }
    L1 ={x| x∈ C1∧x. count≥N×min_sup}
    Return L1
```

Figure 1 Description of improved Apriori algorithm.

3.2. Analysis of improved Apriori algorithm

Compared with the classical Apriori algorithm, the improved Apriori algorithm has three main advantages:

1) The algorithm only visits the transaction database once. On the basis of setting suitable parameters, the algorithm iterates continuously. The algorithm deletes redundant or inconsistent transaction items one by one, which reduces the size of the whole data gradually and saves a lot of time.

2) The data structure of the algorithm is simple. It mainly connects the maximum frequent item sets by two or two sets, and does not use other complex parameters or functions. It avoids the construction of various complex data models or samples in the running process of the algorithm, thus improving the efficiency.

3) The algorithm sets an original transaction code according to the actual situation. Each item in the database only needs to match the previous transaction code, without matching the original transaction code. In addition, the algorithm only needs to link these frequent item sets twice to get large frequent item sets until the maximum frequent item sets are no longer generated. It avoids repeated scanning and wastes time in the scanning process, thus greatly improving the efficiency.

4. Application of Improved Association Rule Mining in Circulation Reading

4.1. Data preparation

Data mining is a new information processing technology. Its main feature is to extract, transform, analyse and model a large number of business data in the database, and extract the key data for assistant decision-making. Data preparation takes a large proportion of the workload in the whole process of data mining, which is a prerequisite to ensure the success of data mining, including raw data collection, data preprocessing and data conversion.

4.1.1. Data collection

This paper uses the data of the circulation and reading system of the library of Ningbo University of Finance &Economics. The system contains all kinds of information. Taking the table of readers as an example, it contains the ID, name, unit, type and the department of reader.

4.1.2. Data collation and preprocessing

Table 1 Rules between Readers and Book Categories

Post	Consequent	Support degree	Confidence level
English	Undergraduates. Humanities College H319	0.126	0.985
Undergraduates	College of Information and Engineering TP391	0.122	1
College of Arts and Media	Grade 12 boys J238	0.117	1
Undergraduate	College of Economics and Management I247	0.107	0.835
College of Humanities	Undergraduates. English H313	0.162	1
Teachers aged 30-40	Female I247	0.189	0.921
College of Humanities Ordinary readers	Female I210	0.253	0.928
VIP readers	College of Arts and Media I247 J238	0.136	0.94
College of Information and Engineering	Information and Management MajorTP368	0.134	0.849

Taking 20,000 people including students, undergraduates and faculty members of Ningbo University of Finance &Economics as the main data source, this paper studies the relationship between readers and Book categories, so only the reader attribute and the book attribute are needed. After the final screening and sorting, this paper takes the total number of books borrowed in 2017 as an example. A total of 18294 readers borrowed books from the library. After screening, the attributes of 18294 readers are reader ID number, reader identity, unit, department, category of books borrowed and other related attributes. This paper focuses on the relationship between the reader and the book category, so as a constraint condition, we select the rules of reader attributes and book categories from all the generated strong association rules, remove some redundant or boring rules, and get 103 rules. Table 1 gives some strong association rules.

4.2. Analysis and Interpretation of Mining Results

In order to make library decision-makers and relevant departments more intuitively see the rule relationship between readers and large categories of books, this paper adopts visual tool effect and presents rules to managers in the form of graphs. Figure 3 shows some association rules. Two or two links represent the relationship between their attribute values. In addition, the thickness of the links represents the strength of the rules. Through Figure 2, we can clearly see which kind of books readers usually borrow under different conditions.

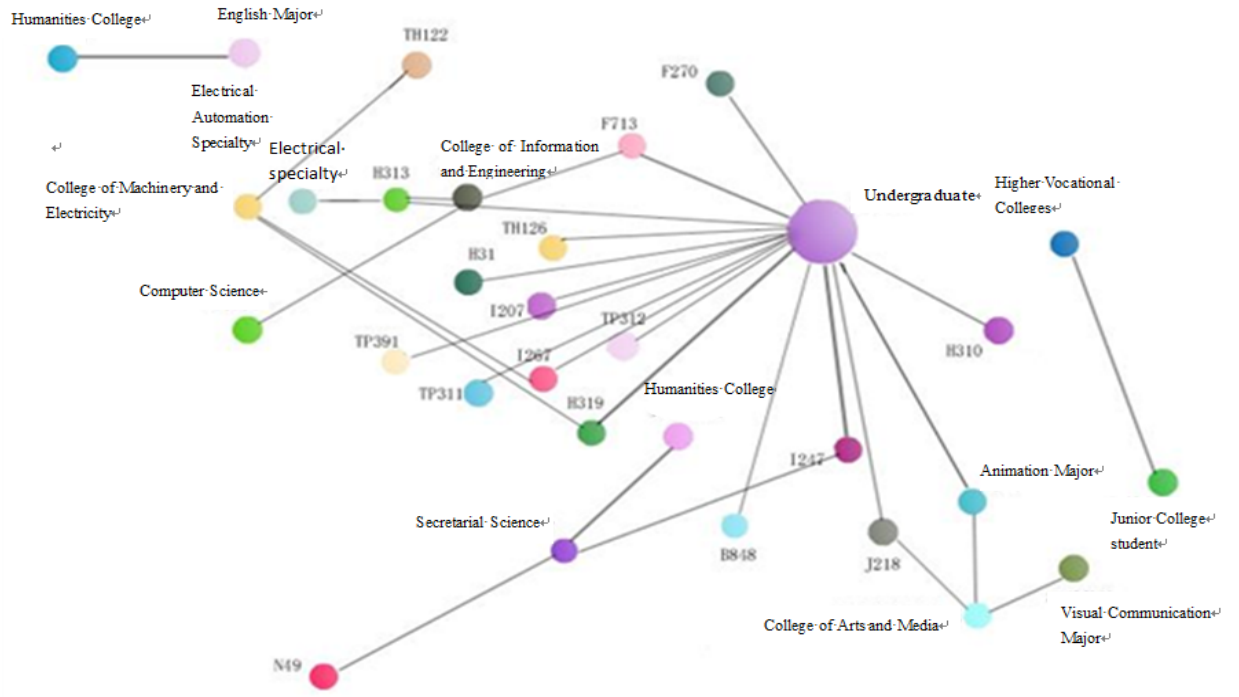

Figure 2 Relevance Strength of Readers and Book Categories

Combining Tables 1 and Figure 2, we can draw the following rules:

(a)Among undergraduates, 0.103 of them borrowed H31 (English) and H313 (Semantic) books at the same time, and 0.92 of them borrowed H31 and H313 books.

(b) The readers who borrow N49 (Natural Science Readings) books are generally secretarial major readers of Humanities College.

(c)Among undergraduate readers, TP393 (computer network category), H313 (semantics, vocabulary, meaning category), H31 (English category), TP391 (information processing category), TP311 (programming, software engineering category), I267 (contemporary prose works category) books are more frequently borrowed.

(d)It can also be seen from the figure that Undergraduate Readers of ICT are closely related to F713 books. In the Institute of information technology, there are many books borrowed from F713 (commodity circulation and market), and a large part of the readers who borrow F713 books are undergraduate students of the Institute of information engineering.

4.3. Decision-making Recommendations

(1) Rational distribution of library collection resources. Putting highly relevant books on the shelves at similar locations (the same floor), such as TH122 (industrial design) and H319 (language teaching), TP393 (computer network) and H313 (semantics, vocabulary, meaning) and TH126 (mechanical drawing) books, saves readers a lot of valuable time and reflects library personality more Humanized service.

(2) The acquisition and cataloguing department shall purchase books according to the proportion of duplicates. For example, class A, class B, class I and class J books have strong correlation, but the borrowing frequency of class A books is relatively small, so the purchase of class A books should be reduced.

(3) Change the traditional bent method. Combining with the actual situation of our library and the needs of readers, we arrange the books with strong relevance on the same floor or in the vicinity. For example, the books of C, D, I and J with strong correlation will be discharged near the location.

5. Conclusions

This paper uses the improved design of Apriori algorithm to analyse all kinds of historical information data of circulation and reading system of Ningbo University of Finance &Economics Library. According to the actual situation, the appropriate parameters of support and confidence are set, and the in-depth mining of circulation data in our library is realized, and more association rules with guiding significance for practical work are obtained. This data mining provides a scientific and reasonable effective method for library information service, information resource collection and scientific arrangement of Library work.

References

[1] Lone, T.A., Khan, R.A. (2014). Data Mining: Competitive Tool to Digital Library. desidoc Journal of Library & Information Technology, 34(5), 401-406.

[2] Nemati, H.R. (2008). Organizational Data Mining: Leveraging Enterprise Data Mining for Management Decisions in Corporate, Special, Digitaland Traditional Libraries .Hershey, PA: Idea Group Publishing, 120-123.

[3] Guo, J.M., Song, S.L., Li, S.L. (2014). An improved algorithm based on Apriori algorithm Computer Engineering and Design, 11, 2814-2815.

[4] Cui, G.X., Li, L. (2013). Research and Improvement of Apriori Algorithms in Association Rule Mining. Computer application, 11, 2952-2958.

[5] Deng, Q.Q. (2011). Library Bibliographic Recommendation Service Based on Disconnected Apriori Algorithms. Library and Information Work, 5, 109-112.

[6] Jia, W.H., Micheline, K.(2003). Data Mining Concept and Data. Machinery Industry Publishing, 23-29.