Data mining applications research based on ISBN management

Yinglan Fang*, Bing Han, Binghui Chen

Department of Computer, North China University of Technology, Jinyuanzhuang Road 5, Beijing, China

Received 1 March 2014, www.cmnt.lv

Abstract

ISBN management systems existed irregularities publishing. How to summarizes the current book publishing rule according to publish information, it can better grasp the overall book market trends based on existing books. This paper introduced data mining ideas to the book publishing field through the study of real-name system to apply business processes. There were two data mining models. One was association rule analysis model based on subject field to book type distribution. It analyzed book type using classic Apriori association rule analysis algorithms to identify book publishing hot and the overall trend of publishing business. It can effectively regulate press book publishing behaviour and has great significant to the China's publishing industry healthy and orderly development.

Keywords: data mining, association rule analysis, ISBN

1 Introduction

The current society is in the period of an information technology, networks, data processing rapid development. How to apply suitable method extract valuable data from the large precious data, it has became an important issue. Data mining techniques emerged. Many universities, research institutions and companies worldwide are turning to data mining and knowledge discovery techniques and gain a lot of researches. So that data mining technology has been rapid development.

This issue research data is come from the ISBN real-name application system. ISBN real-name application system was major project which initiated and promoted by the Chinese General Administration of Press and Publication. It has improved the efficiency and quality of service book publishing industry. Over time, business data among the publishers and audit unit and the General Administration of the system have become rich. As a manager, the Press and Publication Administration eager to get useful part of the existing mass book publishing data to assist them to grasp book publishing direction and standard publishing industry. The research of this issue is to solve the book publishing industry currently facing problems using data mining technology. It would has great importance for the publishing industry.

2 ISBN management data mining analysis

ISBN real-name application system is an reform measures ISBN name to proposed by the Press and Publication Administration in 2006. Its rules is "See draft to get ISBN", "one book only one ISBS," "ISBN real name application." The purpose is to completely change the previous ISBN management and further strengthen publishing management and standardize the behavior of publication. With the accumulation of time, ISBN real-name application system has accumulated one million data, and it had to spend a lot of manpower and resources to maintain these data every year. Managers wanted to get useful information from these chaotic by data some means. Therefore, the data mining in ISBN real-name application system would have a very important significance.

2.1 DATA MINING SUBJECT FIELD

The issue starts from the existing ISBN data items, combining with the relevant provisions of how to optimize and book publishing administration reform publishing behaviour. It determines four sides subject area. They are ISBN distribution usage amount and book type distribution and audience and publishing books standard degrees. ISBN real-name application system's database table name were:

T_B_BOOKINFO (report books basic information table),
T_B_PUBLISHINFO (press basic information table),
T_D_ISBNINFO (audit and pass ISBN data table),
T_D_PLANADD (press add ISBN table),
T_D_PLANTOTAL (press publishers plan to distribute ISBN),
T_D_BOOKINFO_ STATUS (declaration book status table)

and so on nearly 10 table and more than 60 million data items. It creates corresponding data table around the data items needed by selected subject areas and ultimately establishment appropriate data warehouse.

2.2 ASSOCIATION RULE MINING MODEL ANALYSIS

Association rule analysis model is mainly comprehensive

* ***Corresponding author's*** e-mail: jlufangyl@163.com

analysis through the ISBN data items and ISBN real-name application system operation flow. It combines with the basic needs of decision makers, establishes the improved association rule analysis model. it includes selecting data items, filtering and integration, researches and implements Apriori algorithm, selects the appropriate threshold (support, confidence), analysis the resulting data and obtains corresponding conclusions or visualization rendering.

The association rule analysis model in this issue is created according to the press publishing book situation. Its purpose is to find the current book publishing rule within a certain period. For example, from the overall analysis 585 publishers, publishing arts (J category) and the history, geography (K category), 100% of publishers will be publish culture, science, education (G category) books. Which found that the country's book publishing trends and characteristics. The books in ISBN apply real-name system are classified according to the classification of Chinese Library Classification from A category to Z category.

- A Category: It represents the Marxism-Leninism, Mao Zedong Thought, Deng Xiaoping Theory and so on category.
- B Category: It represents the philosophy and religion category.
- C Category: It represents the social sciences category.
- D Category: It represents the political and legal category.
- E Category: It represents the military category.
- F Category: It represents the economic category.
- G Category: It represents the culture, science, sport and education category.
- H Category: It represents the language and writing category.
- I Category: It represents the literature category.
- J Category: It represents the art category.
- K Category: It represents the history and geography category.
- N Category: It represents the natural sciences category.
- O Category: It represents the mathematical science and chemistry category.
- P Category: It represents the astronomy and earth science category.
- Q Category: It represents the biological science category.
- R Category: It represents the medicine and health category.
- S Category: It represents the agricultural sciences category.
- T Category: It represents the industrial technology category.
- U Category: It represents the transportation category.
- V Category: It represents the aerospace category.
- X Category: It represents the environmental science and safety science category.
- Z Category: It represents the general books category.

Due to the every press's basic conditions is different, the different publishing book ability or some other causes leads to the published books type which is scattered. Some rare and poor sale books have small publishing amounts. Some popular and highest-selling books have high issue.

So only from single book publishing number and ISBN circulation amount, it cannot be seen in the overall trend and found publish association rule. Research data mining association rules model and data mining process is the same. They are divided into three processes: data extraction, data processing, association rule analysis.

The first step is data extraction. According to the selected subject areas, the book publishing type, ISBN real-name application system classify the book type, the basic situation of each country published book publishers can be obtained from the database system. It statistically analysis books information resource library and books type distribution of each publishing press. Combined with the actual needs, it select the appropriate data entry according to the association rules and create database table structure. Because it aims at the country's 585 presses and for nearly one million book information in six years, it can extract database in accordance with the certain period of time and different database tables corresponding data. based on these reasons it has created such a table structure: book the original information sheet named Original Book Type (include book serial number, ISBN number, book title, author, introduction, book type, publishers ID, extraction time), so it can prepare data items for the next step of data pre-processing.

The next step is data pre-processing. Book raw information obtained from last step data extraction, it needs for further processing and final processing to these data so as to obtain experimental data of books type distribution and use for analysis of Apriori association rules algorithm. Association rules was analyzed through books type distribution. It extracts book data according to the press group and statistics for each publishing house within a certain period of time books published by type. The statistics result was stored in the statistics books Information table named Book Type Count. Every presses are handled as a single transaction. Each transaction has a number of data items. Books type in the table is the statistics of the total book type within a certain period of time. Association rules algorithm requires incoming data is Boolean variable with the value 0 or 1. That the book publishing house types need to represent with value 0 or 1. These data distribution directly affect the results of association rules, so when it sets this threshold it needs repeated testing to find the most appropriate threshold and compare the various book type statistics value with the threshold. When the book type value is larger than the threshold value it is set to 1 and otherwise its value is set to 0. The handled books is stored to book conversion table named Book Type Static according to its rule. Association rules analysis data table structure is shown in Figure 1:

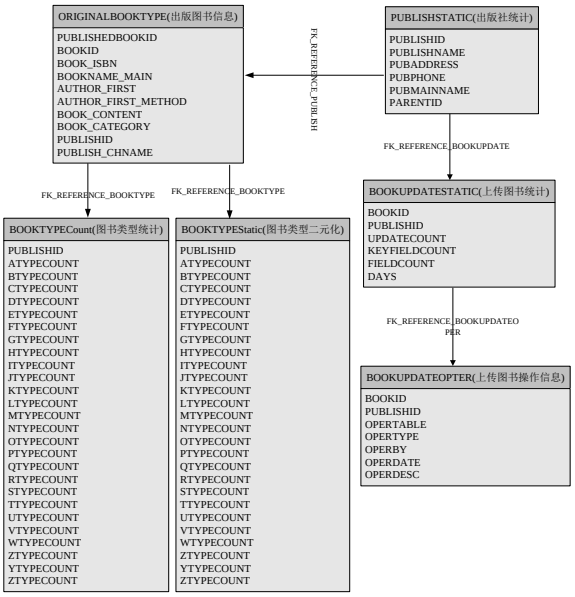

FIGURE 1 Association rules Analysis Database Design E-R

The next step is the results and analysis of association rules. There are two sides. The first side is the parameter setting in Apriori algorithm that is support and confidence threshold. The quality of selected value directly affects the accuracy of the association rules analysis, therefore it requires the combination of the actual situation and the repeated sampling test to help determine the final threshold. The second side is the result of the analysis. It derived some abstract data information after data mining. Researchers and professionals can clearly understand the meaning of the data represents. For general managers it need for appropriate process and convert the data so they can easily accept and understand with text presentation or graphic presentation, and ultimately achieve the purpose of data mining.

3 ISBN data mining design

It can be seen from the research background that the applicable targets to ISBN manage mining system is press administrative department and ISBN issuing departments, including the press and publication administration departments, barcode centres and 31 provincial bureau issued centres. Users can log in the system accordance with their own username and password and use their corresponding function. They can statistically classification analysis and supervise to published books and ISBN usage. It include four aspects. The first is the overall trend in publishing books to current national book press and timely tracking and analysis to book publishing rule. So it can adjustment and control to prevent overheating of some vulgar books published and some less famous books publishing phenolmenon occurs. The second is to supervise the whole process from the application publishing a book to publishing a book so as to discover problematic Press and eradicate illegal acts as "One multi-use", "trading ISBN" and so on. The third is manage publishers sample information cannot be uploaded books by press and the pending book cannot be audited by department. The fourth is the detailed statistical analysis to national book publishing situation. It divides from the time dimension, spatial dimensions and regional dimensions in order to understand the current overall situation of the book publishing industry.

3.1 SYSTEM NETWORK STRUCTURE DESIGN

The system uses B/S architecture model. The user can issue through the browser to many distributed servers on the network request to complete the corresponding business functions. The basic topology of the network structure is shown in Figure 2:

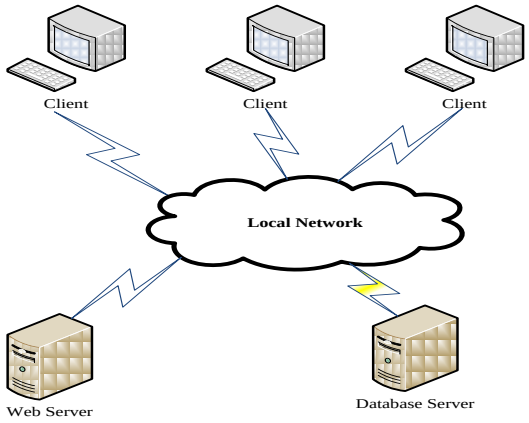

FIGURE 2 ISBN information data mining system network topology

Interaction between the user and the system are described as the below:

- User fills out the network address of the system in the browser and request a service to a Web server on the network via a Web browser. Web server user identity verification after transferring to the desired page using the HTTP protocol client and displayed in a Web browser.
- Web servers accept user requests and connect the database. Then put forward data processing request to the database server through SQL and transfer the results to the Web server.
- Web servers obtains visualization results from data service processing results and returns to the browser user.

3.2 SYSTEM MAIN CLASS DESIGN

In this paper, it has designed the interfaces and classes in association rules analysis algorithm (Apriori algorithm). Apriori algorithm is mainly read transaction data in the data warehouse by Get-Source method. Then through combination of internal processing transaction data, it fits

the data structure in line with the main function, and ultimately get frequent data mining item sets. The Apriori algorithm class design of association rules analysis are as follows (Table 1):

TABLE 1 Apriori Test class method function

Class name	method name	function
AprioriTest	Public String getRules ()	get related combination of frequent items
	Public Integer getconfidence ()	get a given minimum confidence
	Public Integer getSupport ()	get a given minimum support
	Public void mine ()	mining each frequent item sets
	Public void associationRule ()	mining frequent association rules
GetSource	getBookTypeCount ()	get the book information to data type
	getBookTypeStastic ()	get book type information
	getSet ()	get association rule analysis datasets

4 ISBN Data Mining Implementation

This section describes the main functions of association rules analysis modules. It includes raw data extraction, data conversion, data filtering, association rules analysis and graphical display.

FIGURE 3 Raw data statistics Schematic of association rules analysis

4.1 APRIORI ALGORITHM FOR DATA PRE-PROCESSING

1) Data statistics. It statistics and classifies source data extracted from the previous step and classifies according to classify method. Specific statistical results are shown in Figure 4.
2) Data filtering and data cleaning. It further processes data counted by the previous step, converts into binary data structure representation by a given threshold and provides analytical data for the association rule algorithm. It is as shown in Figure 5:

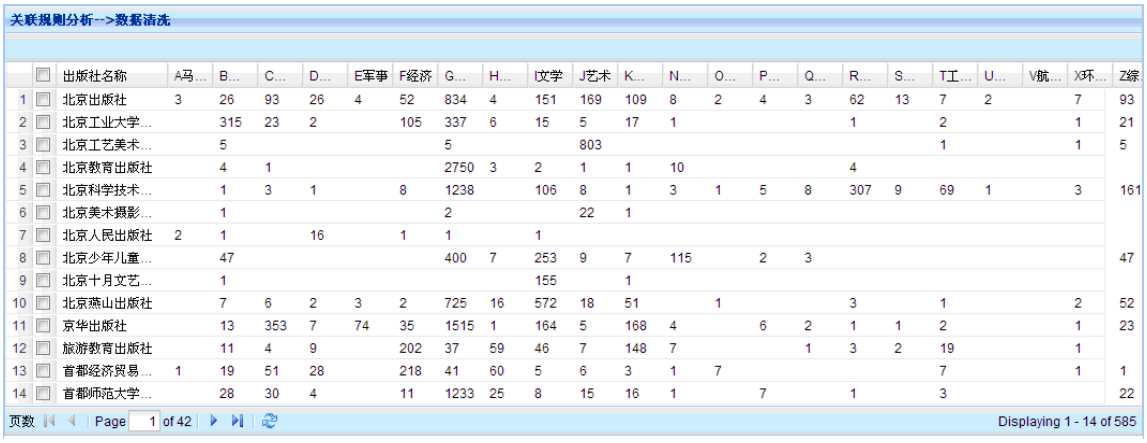

FIGURE 4 Association rules analysis data filtering schematic

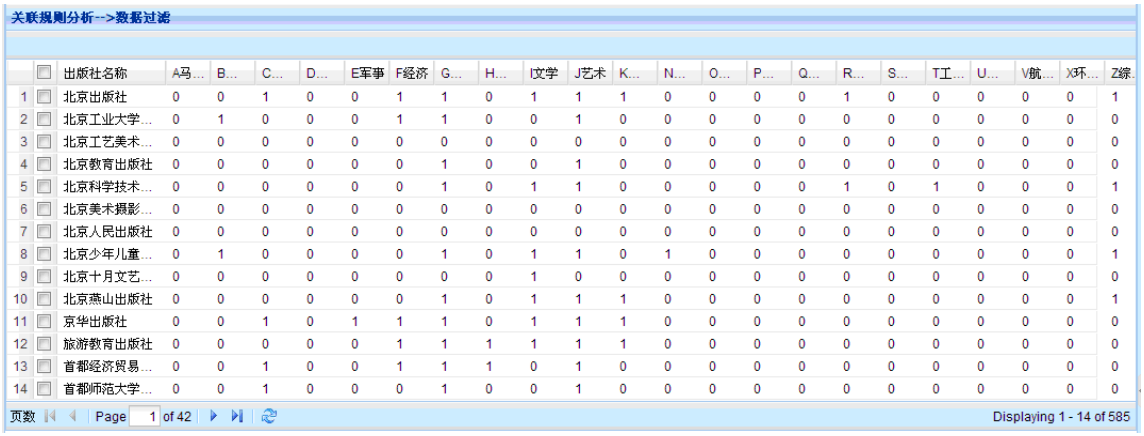

FIGURE 5 Data conversion schematic in association rules analysis

4.2 APRIORI ALGORITHM DATA ASSOCIATION ANALYSIS

The purpose of this analysis is the association rule 585 publishers across the country to conduct the type of book publishing association rules analysis, which found that the association between a certain period of time and the overall trend of analyzing summarizes book during this time.

It can be seen from the Figure 5 that it has 633,591 books and 585 press, namely the number of transaction in association rules analysis process is 585. In line with the specified degree of support minSupport equal to 0% and confidence minConfidence equal to 30%, there are frequent item sets conditions. Namely: {[I, J] =>G}, {K=> [G, J]}, {[G, K] => J}, {I => [G, J]}, {G =>J}, {[J, K] => G}, {[G, I] => J}, {J => G} which rules is [I, J] => G and confidence levelis 48% and degree of support is 100%.

It illustrates that 100% nationwide press publishes science books and economic books and culture, science, education, sports books. Among 48 percent press which published class I (social sciences) and J books (economy class) may also publish books of the class G (culture, science, education, sports).

From an overall perspective, we can see that the economic books, culture, science, education, sports books, literature books almost become mainstream during this time.

关联规则集	置信度	支持度
I,J => G	0.47692308	1.0
K => G,J	0.36923078	0.875
G,K => J	0.33333334	0.9692308
I => G,J	0.53675216	0.91719747
G => J	0.53675216	0.91719747
J,K => G	0.32307693	1.0
G,I => J	0.4923077	0.96875
J => G	0.774359	1.0

FIGURE 6 Association structure data shows in association rules analysis

4.3 ASSOCIATION RULES ANALYSIS RESULTS SHOW

The results association rules analysis is a string of characters and numbers. For which the system user is difficult to understand the meaning. It can be very clearly seen through the graphical display that the roughly distribution of books types and can help verify the accuracy results information of the association rules analysis (Figure 6-7).

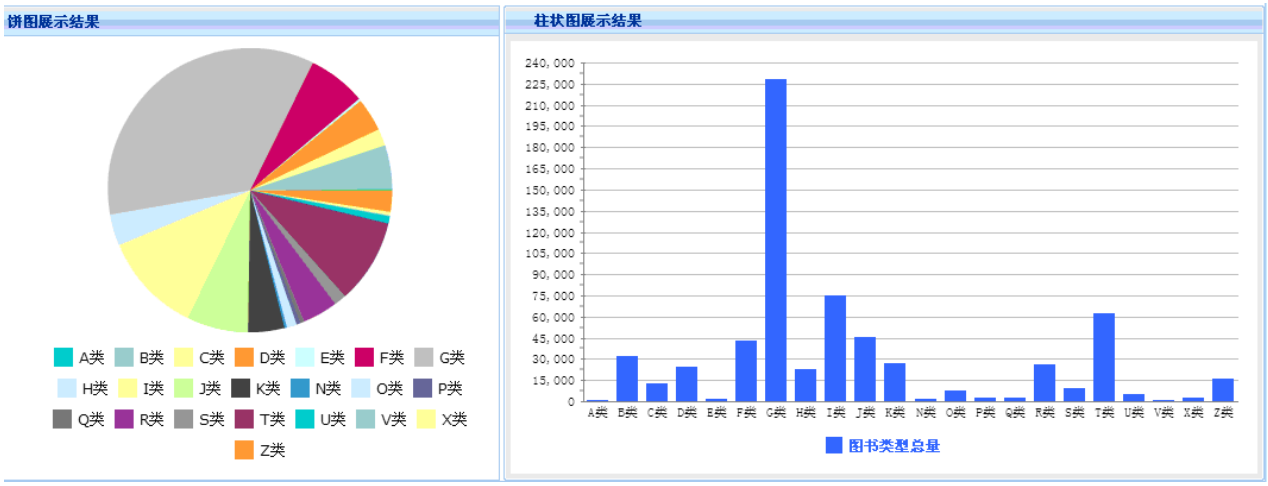

FIGURE 7 Graphical association rules analysis shows

5 Conclusions

In this paper, data mining technology has been imported into the book publishing industry, and conducted in-depth theoretical study and specific mining algorithm experiments. This paper mainly studied the association rules analysis Apriori algorithm. According to the needed analysis problems and actual application, it deeply sees the business of the ISBN real-name application system. Combined with existing data mining techniques, it has analyzed and created around the theme of type press books domain analysis model and association rules around behaviour publishing books degree of clustering model specification subject areas. From different angles ISBN statistical analysis of the data item information, the evaluation and mining results graphically is shown.

Acknowledgments

The work in this paper has been supported by Beijing Natural Science Foundation (No. 4131001). It is also supported by National Science and Technology Support program (No. 2012BAH04F00) and Beijing Natural Science Foundation (No. 4132026) and Beijing Education Science and Technology Development plan (No.KM 201210009006) and Beijing Image Information Processing and Intelligent Recognition Scientific Research Platform Construction (No.PXM2012-014212-000024).

References

[57] Liu Y 2006 Data mining in the management decisions of college students use pattern analysis *Chengdu University of Information Technology* **21**(3) 373-7 *(in Chinese)*
[58] Wu X 2009 Data mining Application of in Educational Administration technology *Harbin Engineering University master's thesis* 16-7 *(in Chinese)*
[59] Cai X 2007 Telecommunications data mining data preparation process standardized design *Computer Engineering* **33**(24) 44-5 *(in Chinese)*
[60] Jun X 2012 Data mining Application in the stock market technical analysis *Zhejiang University Thesis* 19-21 *(in Chinese)*
[61] Cheng J 2011 Research and analysis algorithms of multiple association rule mining *Changchun University of Technology (Natural Science)* 107-9 *(in Chinese)*
[62] Wei J, He Pi, Sun Y 2005 study text hierarchical K-Means clustering algorithm *Computer Applications* 23234 *(in Chinese)*
[63] Xia N, Su Y, Qin X 2010 An efficient k-medoids clustering algorithm *Application Research of Computers* 45178 *(in Chinese)*
[64] Li Q, Yuan J 2012 A text clustering algorithm based on optimal density DBSCA *Computer Engineering and Design* 1409-10 *(in Chinese)*

Authors

Yinglan Fang, born in August 1973, Shangcheng Henan, China.

Current position, grades: associate professor and master supervisor at North China University of Technology.
University studies: ME degree in Computer System Structure at Jinlin University of China in2002.
Scientific interests: computer application and computer security.
Publications: 12 papers.

Bing Han, born in August 1971, Jiaozuo Henan, China.

Current position, grades: associate researcher and master supervisor in North China University of Technology.
University studies: ME degree in Computer System Structure at Jinlin University of China in 2001.
Scientific interests: software engineering and computer control.
Publications: 2 patents, 13 papers.

Binghui Han, born in September 1987, Luoyang Henan, China.

University studies: ME degree in Computer Technology at North China University of Technology in 2013.
Scientific interests: software engineering and data mining.
Publications: 1 papers.