The application of OLAP and Data mining technology in the analysis of book lending

Xiao-Han Zhou[1,a], Xiao-Mei Zhang[2]

[1]Library of Southwest University, Chongqing 400715,China;
[2]ChangJiang Chong Qing Waterway Bureau, Chongqing 400715,China;
[a]lichking@swu.edu.cn

Keywords: Book lending, Data warehouse, Cubes, Association rule.

Abstract. Book lending data is the core of the library's basic business, Through the use of OLAP and data mining technology, the library can be effectively mine the accumulated library books lending data, provide the basis for the development and daily management of the library. This paper realizes the design of data warehouse based on the book lending data of the library, On the basis of this, a cubes is established, through a number of dimensions to analysis the book lending data. Then the GRI algorithm based on data mining is used to mine the data effectively.

1 Introduction

Database technology is developing rapidly in the application of the library, readers in the library to borrow books at the same time, also produced a large number of books to borrow data. Some experts pointed out: " By using mining analysis method of the association rules on the circulation data of a certain reader group in a certain period of time, can found the implicit association between the knowledge of different subjects when readers in professional learning, this has an important guiding significance for the library to adjust the structure of resources and improve the level of service for readers."[1] Through re-analysis and re-prediction on a large number of readers and reading behavior, we can evaluate the effect of the structure of the collection structure, and make the corresponding adjustment. [2]

Many foreign libraries, data mining technology has been a certain application. This paper based on the book lending data of a college library for many years in domestic, using data warehouse technology to organize and concentrate the data effectively, then use online analysis OLAP and data mining two kinds of intelligent technology for data mining analysis.

2 System overall framework

Data analysis of book reading need to be built on the overall framework of the book lending data warehouse, this thought and theory of data warehouse was introduced in 1993 by W.H.Inmon.[3] By building a data warehouse of library book lending, the research object is the evaluation of the data analysis of the book lending, demand analysis, data warehouse concept model, logical model and physical model design for the data analysis and evaluation of the book lending are emphatically described, and do data preprocess before the establishment of data warehouse, store the standardized and unified data in the data warehouse. The overall framework of the entire book lending data analysis system is shown in figure 1.

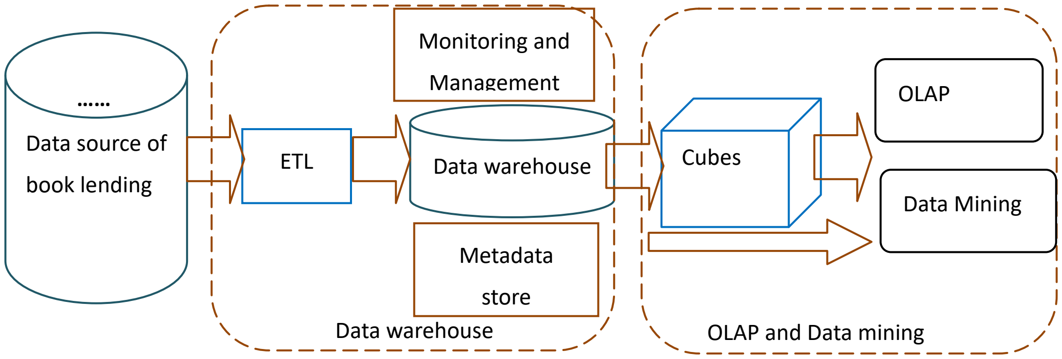


Figure 1: The overall Framework of the entire book lending data analysis system.

The system is mainly composed of three parts: Data source of book lending, Data warehouse, OLAP and Data mining. Among them:

1, Data source of book lending. University library services mainly for teachers and students to lend books, the system data because of policy adjustment or relocation and other objective reasons are not part of the standard data, therefore, this paper main selected the book lending data of a university library from 2010 to 2015 years.

2,Data warehouse. Book lending data first through the data extraction, conversion, loading and other ETL process, that is, data cleaning, according to the requirements of a unified format integrated into the data warehouse. Among them, Data warehouse is responsible for storage analysis, decision data; monitoring and management is responsible for the management of data warehouse; metadata storage is responsible for the management of metadata.

3, OLAP and Data mining. After the Data warehouse is built, the cube is built on the basis of the Data warehouse. Using cube can carry out OLAP multidimensional analysis, also can carry on data mining, data mining can directly mining detail data of the data warehouse. Among them, the OLAP application refers to the multidimensional analysis of OLAP and the front end display of OLAP, which is the report tool, query tool and data analysis tool. Data mining involves pattern evaluation, and the results are displayed in a tabular or graphical form [4].

3 Establishment of Data warehouse

There are two ways to design a data warehouse: top-down and bottom-up. The bottom-up design method emphasizes the application decision data, according to the requirements of the application to obtain data, design dimensions and cube is the beginning of the project. The bottom-up design method firstly creates the data warehouse according to the requirement analysis of the system, and then sets up cube according to the different analysis topic. After the analysis of the demand of the book lending data and the book lending relevance, in this paper, we use the bottom-up design method to build a data warehouse based on the relationship between the book lending data.

3.1 Logical model and physical model construction of Data warehouse

At present, there are 3 main types of data warehouse multidimensional data model: star schema、snow flake schema、galaxy schema. Compared to the star schema and snow flake schema, the galaxy schema has a higher accuracy of data, this paper uses the galaxy schema. Galaxy schema is more than one fact table sharing one or more dimensions tables, this data model can be divided into two types: the fact data table and the dimension data table.

The fact data table is used to store the fact data of the data warehouse, and is the largest table in the data warehouse. The fact table contains the details of the book lending data. The dimension data table is used to store the dimension data of the data warehouse, and the observation of the dimension data table is the basic driving force of business intelligence. This data warehouse design mainly has the following four kinds of dimensional data sheet: the reader dimension, the collection dimension, the book dimension, the time dimension. Figure 2 is the subject schema of the book lending data.

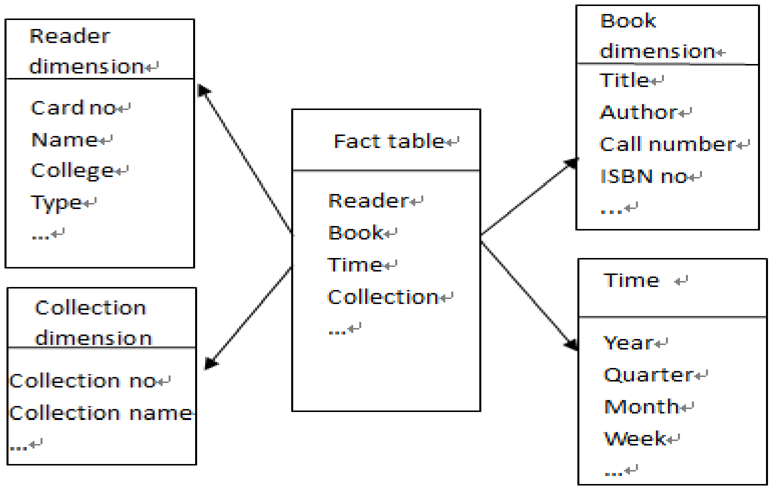


Figure2: Subject schema of the book lending data.

After determining the data table and the dimension data of the logical model of the data warehouse, we need to determine the physical model. The physical model of data warehouse is the realization mode of the logical model of the data warehouse in the physical system, which including a variety of physical tables in the logic model of the specific. For example, data structure type of the table, index strategy, data storage location and data storage allocation and optimization operation of physical model, etc.

3.2 Data extraction, transformation and loading

After the design of data warehouse model, we need to use the ETL tool to load data into the data warehouse. ETL included data extraction, data transformation, data loading, which is responsible for the completion of data from the source data to the target data warehouse conversion process, is an important step to create a fact data warehouse, and also is an important part of building a data warehouse. Users extract the required data from the data source, after data cleaning, and finally loaded into the data warehouse according to a predefined good data warehouse. In the ETL process, this paper uses SSIS data flow components, including search, fuzzy search, fuzzy grouping, data conversion, dimension conversion, the script component, OLE DB commands, and conditional Split etc.

4 OLAP and Data mining

4.1 OLAP analysis

OLAP multidimensional analysis based on the cube, a cube is a data view of user for data analysis, which is a data model for analysis, it can provide a variety of observation angle of view and analysis oriented operation to analyst. By using Microsoft SQL Server 2008 Analysis Services to create a cube, you can browse through and analysis the data of the cube's. The action of multidimensional analysis included: drill, slicing, dicing and rotation.

Drilling operation is the operation of the data from the depth of the user, in the multi-layer data through the drilling operation allows the user to view the data at different depth levels. The drill-down is thorough observation, can see deeper data refinement; the drill-up is generalize, can summarize the details of the data to high-level. For example, by drill-down and drill-up operation can get annual, quarterly and monthly book circulation. According to the amount of lending we can generate the required graphics statements, as shown in figure 3.

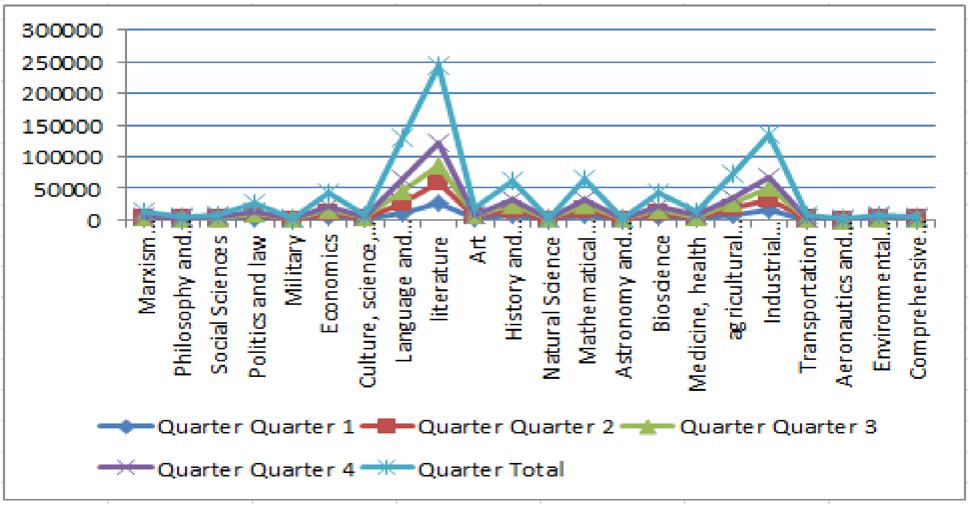


Figure3: Lending data of all kinds of books in the quarter of 2010.

The object of the slice operation is a dimension of the data set, which is obtained by a plane data. Such as the analysis of the way of reading the book is slice operation in the dimension plane of the way of reading. There is also can analyze the lending data of the collections from the collection dimension, which is also a slice operation. Cutting operation is a multidimensional operation, it can help to view the data combined form multiple angles. In short, to meet the analysis demand of library staff, need to observe data from different levels, different depth, you need to drill, cut and other operations to analyze cube continuously, in order to achieve the purpose of finding information.

4.2 Data mining

Data mining is very important to find and describe the hidden patterns of the specific cube. SQL SERVER 2008 provides decision tree, clustering analysis, association rule algorithm and so on for data mining, help the decision-maker automatic searching mode and interactive analysis, solve the cube data quantity increase fast but difficult to find hidden information problems.

In this paper, the relationship model between the readers to borrow books are established, the purpose is to excavate the relationship between the readers and the books. For example, whether there is a certain link between borrow "literary books" and borrow "economics books". Therefore, the Generalized Rule induction algorithm is used. The main application object of GRI algorithm is transaction database, and the GRI mining of a transaction database can be described as follows:

Set $I = \{i_1, i_2, \ldots, i_n\}$ be a collection of different items of n, Each $i_m (m = 1, 2, \ldots, n)$ is called a data item, transaction T is a subset of data set I. Each transaction has a unique identifier that is connected to the T_{id}, the whole transaction set D is composed of all the different transactions.

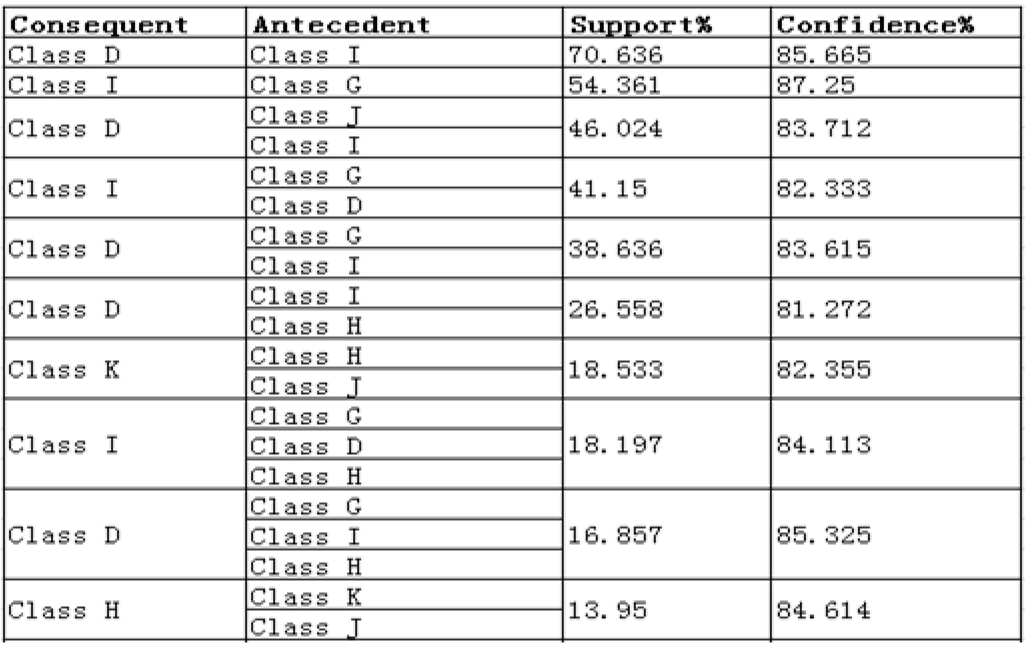
Set $X \subseteq I$ be a transaction item set, B is the number of transactions X that are contained in the transaction set D, A is the number of all transactions that are contained in the transaction set D, then the support of the data item set X is defined as: $Support(X) = B \mid A$

An association rule defined on I and D like $X \Rightarrow Y$ is given by satisfying a certain confidence. The so-called rule of the confidence refers to the number of transactions containing X and Y and the ratio of the number of transactions containing X, among $X \cap Y = \phi$, $X \cap Y = \phi$.

$$Confidence(X \Rightarrow Y) = Support(X \cup Y) \mid Support(X)$$

Given a transaction database, association rule mining problem is to find the appropriate association rules by the user specified minimum support and the minimum confidence. Association rule mining firstly finds out the frequent item sets with the minimum support degree that is greater than or equal to the user in the transaction database. Then use the frequent item set to produce all the association rules, according to the user set the minimum confidence to make a choice, and finally get the strong association rules.

By using GRI algorithm mining library lending data, such as class D (political, legal), class F (economic), class G (culture, science, education and sports), class H (Language and Literature), class I (Literature), class J (Art), class K (history, geography). Book lending relationship model to see the following chart:



Consequent	Antecedent	Support%	Confidence%
Class D	Class I	70.636	85.665
Class I	Class G	54.361	87.25
Class D	Class J Class I	46.024	83.712
Class I	Class G Class D	41.15	82.333
Class D	Class G Class I	38.636	83.615
Class D	Class I Class H	26.558	81.272
Class K	Class H Class J	18.533	82.355
Class I	Class G Class D Class H	18.197	84.113
Class D	Class G Class I Class H	16.857	85.325
Class H	Class K Class J	13.95	84.614

Figure4: Book lending relationship model.

Thus, there is a certain relationship between readers to borrow books, such as lending class I books readers, more likely will borrow class D books; borrow G and I class library readers, the possibility of a larger will borrow D books; Through the use of the results of the data mining, when the library to recommend books to readers, it can be more targeted recommendations.

Acknowledgments

This work was financially supported by Southwest University Fundamental Research special funds project (XDJK2014L077).

Reference

[1]Yuhui Wei, Jie Pan. Quantitative analysis method of book circulation data association mining [J] modern information, 2005,11.108-110.

[2] Nanqiang Xia, Hongmei Zhang. Digital library personalized service based on data mining [J] library science research, 2006,1.32-34.

[3]W.H.Inmon, Building the Data Warehouse, Chapter 2,1993.

[4]Hongping Fang, design and implementation of OLAP system based on data warehouse. Journal of Wuhan University of Science and Technology, 2004, 27 (1): 69-71.

[5]TANG Z H,MACCLENNAN J.data mining principle and application of -SQL Server 2005 database [M]. Fang Zhu Kuang, Xianlong Jiao, et al. Beijing: Tsinghua University press, 2007.277-287.