THE INTEGRATION SYSTEM FOR LIBRARIANS' BIBLIOMINING

SHIEH, JIANN-CHERNG
Graduate Institute of Library and Information Studies, National Taiwan Normal University, No.162, Sec. 1, Ho-Ping E. Rd., Taipei, Taiwan
jcshieh@ntnu.edu.tw

Introduction. Over the past decade, data mining has been widely applied for specified purposes in various institutions. For library services, bibliomining is concisely defined as the data mining techniques used to extract patterns of behavior-based artifacts from library systems. Bibliomining process includes identifying topic, creating data warehouse, refining data, exploring data and evaluating results. The cases of practical implementations and applications in different areas have proved that the properly enough and consolidated data warehouse is the critical promise to success data mining applications. However, the task to create the data warehouse is highly database techniques dependent and involves much information engineering knowledge. It certainly hampers librarians, who were not trained in database discipline and are with little database literacy, to apply data mining technique to improve their work flexibly. Moreover, most marketed data mining tools are even more complex for librarians to adopt bibliomining in library services and operations.
Method. We apply rapid prototyping software development method to develop the integration system. Those who joined the developing procedure are one database designer, three librarians, one system analyst, two library domain knowledge experts and one programmer. The system is designed based on library experts' views and librarian capability on their domain knowledge.
Results. We propose a bibliomining application model and have developed an integration system for librarians' bibliomining in easy and flexible usage of library data mining operations.
Conclusion. The primary job of bibliomining is to discover what meaningful and useful information to aid decision makings for library managers. They must pay much attention on how to meet their requirements. The developed bibliomining integration system meets the purpose and can help librarians do data mining works well.

Introduction

Having attracted a lot of attentions recently, data mining is a new technology to tackle new problems with great potential for valuable discoveries in various application fields. Data mining is the process of extracting meaningful or useful patterns and rules from large data sets or huge databases. Many of most successful applications of data mining are in the marketing and customer relationship management areas. Through data mining detailed behavioral data on existing customers culled from operational systems, enterprises hope to turn these myriad records into some sort of coherent profile of their customers in order to improve the quality of services.

Bibliomining, or data mining for libraries, is the application of data mining and bibliometric tools to data produced from library services [2] [8] [9]. In order to meet the needs of different patron groups, libraries can apply data mining process to uncover patterns or rules of artifacts of use in communities that gain library services [3] [7]. Furthermore, data mining can also be applied to discover effective information from library operation-related data sources to aid in the support of library management decision-makings [1] [3] [4] [5] [6] [10] [11].

In the process of data mining or bibliomining, we need to do the work of data extraction and transformation from required data sources to have a clean and available data warehouse or regular base. It is obviously to see that the properly enough and consolidated data warehouse is the critical promise to success data mining applications. For specified purposes, some current ETL (Extraction, Transformation and Loading) tools help database experts do the job well. However, such highly database techniques dependent task, involving much computer and information engineering knowledge, hampers librarians to use the bibliomining to improve their work flexibly. Thus, the friendly user interface for bibliomining is much important and required to librarian usage.

In this paper, we propose a bibliomining application model and develop a corresponding prototype of integrating data sources based on library expert views and of providing user-friendly bibliomining interface based on librarian requirements. Under the model, we expect to get rid of the complications to advance data mining in library services and management.

Bibliomining in Libraries

Data Mining

Data mining offers powerful and effective techniques for uncovering useful and meaningful information in voluminous datasets. It has been used successfully in many communities for tracking behaviors of individuals and groups. Data mining is an interactive process that typically involves four phases as illustrated in Figure1: problem definition, data preparation/extraction, modeling/evaluation and presentation.

Domain managers initiate the data mining objectives. Data mining experts and domain experts work closely together to define the problems and the requirements from the objective perspectives. It is important to verify that the data meets the requirements for solving the identified problem. Domain experts understand the meaning of the data. They collect, describe, and explore the data to build the data model. Then, from various data sources, data miming experts or database experts prepare the data for the model by extracting tables, records, and attributes, cleansing and formatting the data and also creating new derived attributes using ETL tools. The modeling and the evaluation are coupled. Data mining experts select and apply different data mining functions with changing parameters for the data warehouse until to get optimal values. Frequent used mining functions are clustering, association, classification, sequencing, outlier and regression. Finally, we present the results by using visualization tools and then implement the results.

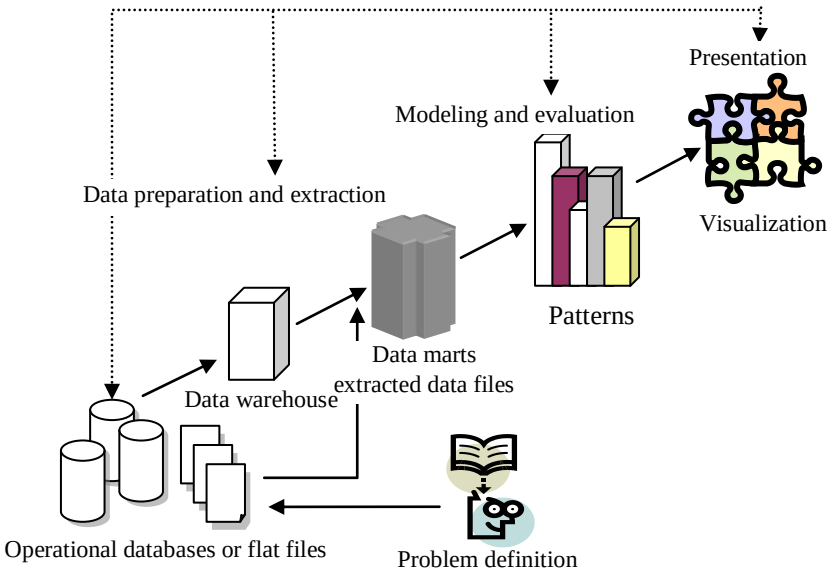

Figure 1. The process of data mining.

Bibliomining

Data mining techniques can help libraries in knowing the trends of popular subjects to enable better focus of acquisitions and budgets, analysis of usage, borrowing and interlibrary loan patterns to plan collection, time-of-day traffic to plan opening hours and staffing, etc. The goal of much data mining focuses on patron services to have better marketing, personalization and targeted collections.

Bibliomining is the application of data mining and bibliometric tools to data produced from library services to aid decision-making and justify services. The term bibliomining was first used by Nicholson and Stanton in discussing data mining for libraries. The bibliomining process consists of: determining areas of focus; identifying internal and external data sources; collecting, cleaning, and anonym zing the data into a data warehouse; selecting appropriate analysis tools; discovery of patterns through data mining and creation of reports with traditional analytical tools; and analyzing and implementing the results. The process is cyclical in nature [8].

Bibliomining is just another synonymy of libraries' data mining that different experts are also involved in the process. Domain experts, or librarians or library specialists, identify required data sources to provide specific services, to resolve management issues or to help decision-makings in library. Data mining experts or database specialists take the responsibility for collecting data, cleaning data and transferring data or building data warehouses. Data mining experts select proper tools to discover meaningful patterns to predict or to describe different patrons or clusters of demographic groups that exhibit certain characteristics. Under the process, the bibliomining application process can be depicted in Figure 2.

For example, librarians would like to improve the efficiency of circulation service. They should know how to arrange manpower on the desk. Library specialists may have idea to analysis the time spans of historical circulations to get helpful information. They thus identify the past 5 academic years' borrowing records and patron data as required data resource for data mining. Data mining experts or database specialists then begin to dig out the required historical circulation data from backup data repositories of S and I library automation systems. However, the formats or types of fields used to record circulation data are not quite equal completely, such as 10 characters for I's student ID but 8 for S's. Data mining experts or database specialists must do data conversion by SQL programming to resolve the case. After required data are well-cooked into SQL server database, according to their experiences on mining similar cases, data mining experts or database specialists apply association algorithm to mine the database. Then we will have the results if we give much more prayer. Until the resulted information is verified and proved by library specialists and librarians, and the library administrators agree them, we finish the bibliomining case.

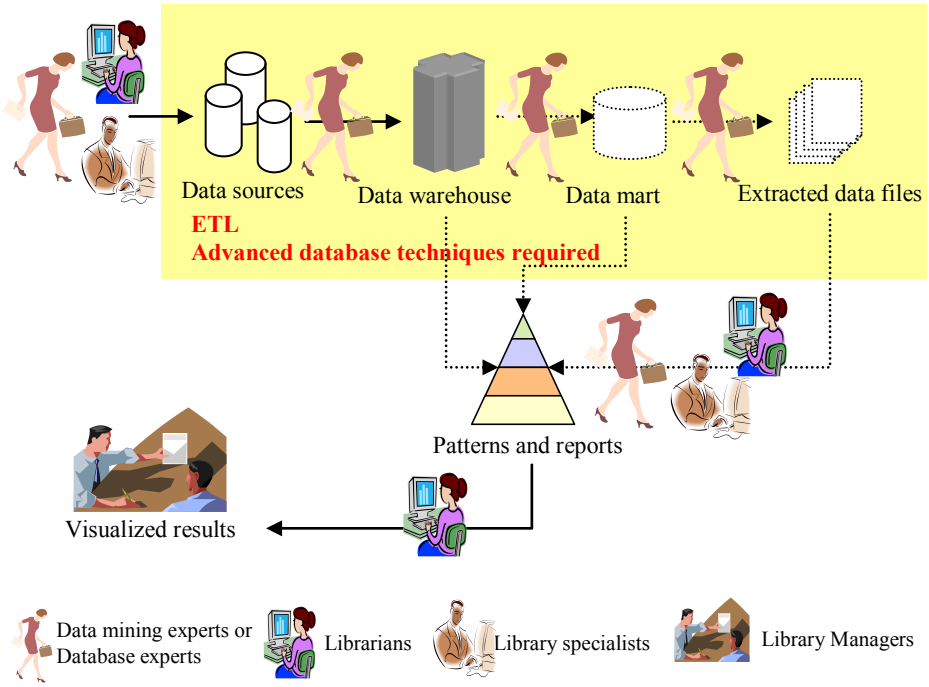

Figure 2. The flow of bibliomining

In practical implementation, the bibliomining requires more database techniques such as SQL programming in data extraction. Thus, librarians would have less confidence on applying bibliomining to solve critical library problems. Even many diverse ETL tools exist, the database specialists are ham-handed to take data collection, extraction, transformation and loading to build initial data warehouses required by data mining case by case. The situation really obstructs flexibility for bibliomining in libraries. In next section, we propose a bibliomining application model and develop a corresponding prototype of integrating data sources based on library experts' views and of providing user-friendly bibliomining interface based on librarians' requirements. Under the model, we expect to get rid of these cumbersome complications to advance data mining in library services and management.

Bibliomining Application Model

What most important for librarians to discover useful service and management information through bibliomining is to have a user friendly interface and even technology free operations. Thus, librarians or library experts can have much attention on critical issues about libraries themselves. The ideal practice model can be described as in Figure 3. In the model, librarians are simple users who just drill down the menus and pick up what related data items they concern. The corresponding data cube is then created and next feed to on-line analytical processing (OLAP) or data mining tools for further processing and presentation.

For data sources integration, data mining experts and library specialists can first define relationships among them comprehensively. Data sources formats can be databases, Excel files and text files. When librarians propose a specific problem, they do drilling and picking operations on needed data items or fields to constitute the problem-solving data cube. Including generalized concept hierarchy, experts can classify, categorize, cluster data as they require. Librarians can be free to database technical operations to generate and present the results..

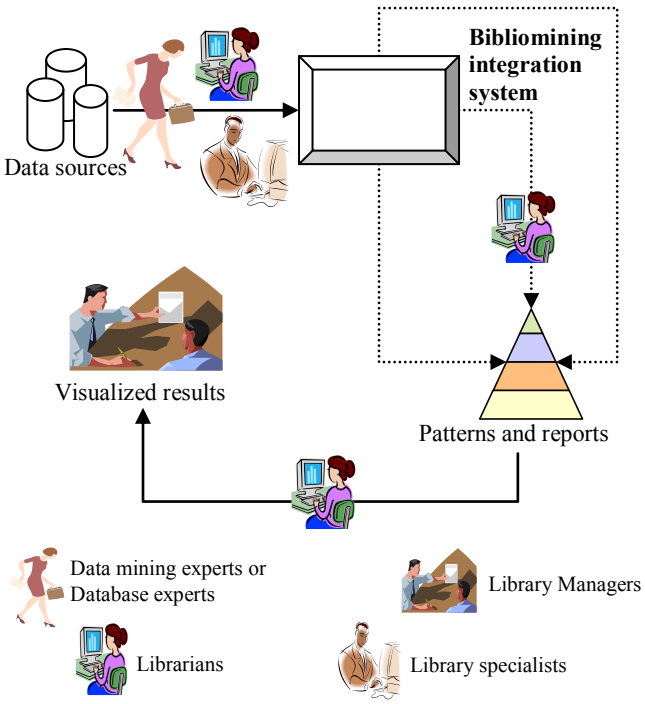

Figure 3. Bibliomining application model

Thus we need the bibliomining integration system to implement the model. The bibliomining integration system mainly provides functions for expert administrations and librarian operators to set various data relationships, to define data concept hierarchy, to select and integrate data items, to

diagram data warehouse model (star schema and snowflake schema), to transform and load data, to create specific cube, to import data to on-line analytical processing or data mining tools, and to output the visual results or patterns. Data mining experts and library specialists have much responsibility of integrating critical and necessary data sources to construct a comprehensive huge data warehouse. Librarians only do the jobs of selecting required data items to create the cube in order to resolve a specific issue. The system structure is depicted as Figure 4:

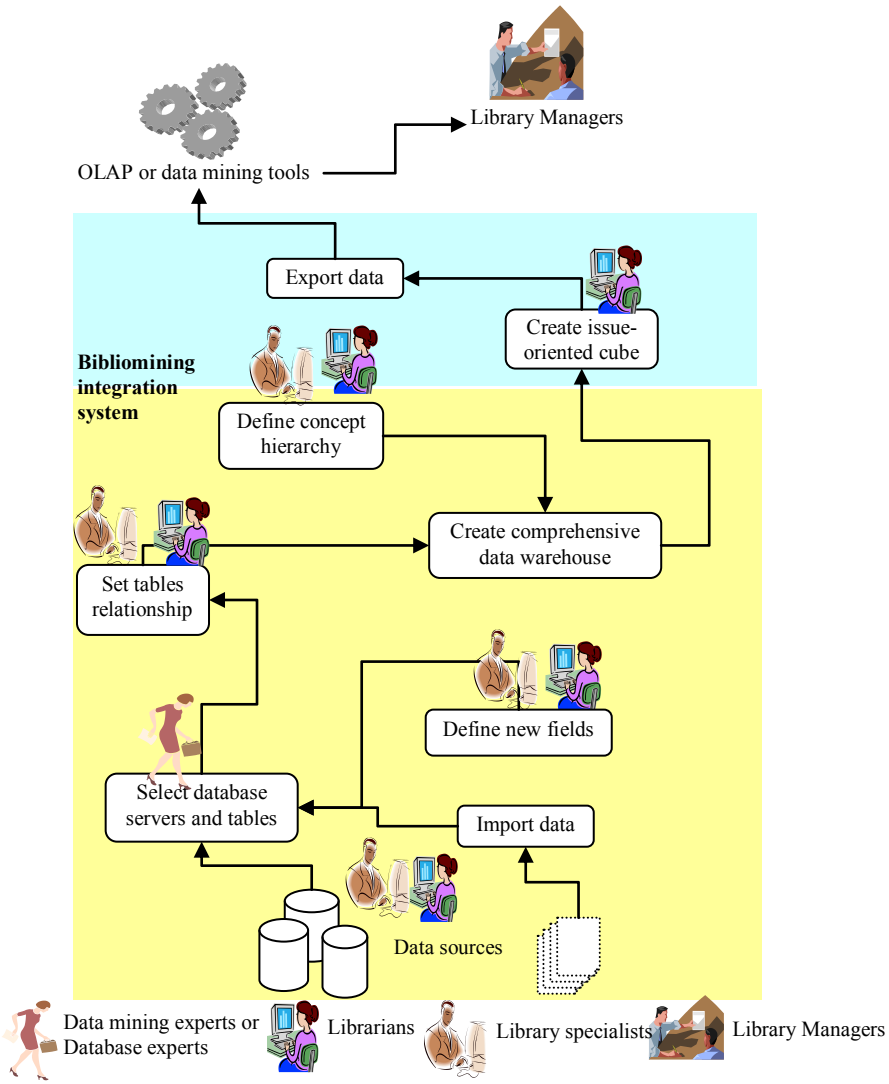

Figure 4. Bibliomining integration system structure

1. The system handles various required data resources such as different databases and data files for library issues.
2. The non-database data can be directly imported from text files to constitute tables.
3. According to requirements, library specialists or librarians can define new data fields, or pick necessary data items or fields to have new tables, and select database tables from different servers to set their relationships.
4. For information analysis purpose, library specialists or librarians can define concept hierarchies of digital, date and character data.
5. The system creates a comprehensive data warehouse and provides for different issues of a specific library.
6. Librarians can create issue-oriented cube to resolve individual problem and to export it to data mining or OLAP tools.

Based on the proposed bibliomining application model, we develop the prototype to implement the integration system. Data sources can be various formats of MS Access, Excel, SQL Server and text. Exporting formats are for MS Excel, SQL Server, and SPSS Clementie. The prototype is built on MS Windows Server 2003 and SQL Server 2005 using MS Visual Basic. Concept hierarchy definition function provides library specialists to define hierarchy of any data field for aggregating data purpose. Data warehouse or data cube are implemented to support star schema and snowflake schema models. Some operating windows of the prototype are shown as follows:

1. Based on the data sources required by library specialists or librarians, data mining and database experts use the functions provided by database management system or SQL programming to select required tables from database servers and set fact table.

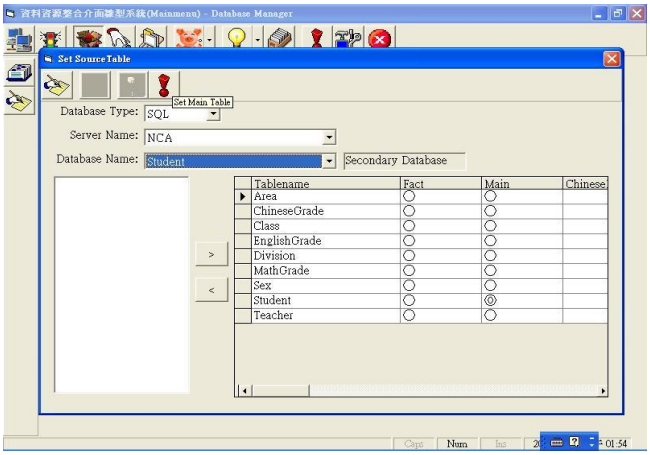

Figure 5. Database server and tables selection

2. According to the picked tables, data items and setting relationships, the integration system generates the snowflake database schema. Currently, the system provides snowflake and star schemas.

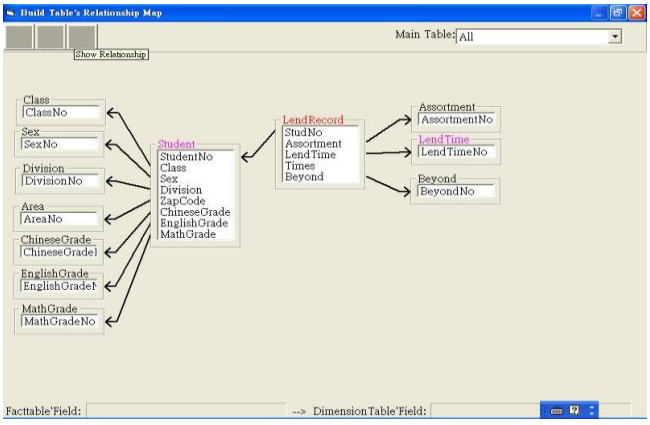

Figure 6. Snowflake schema model building

3. Library specialists or librarians can define concept hierarchy not only for number or date data but also for character data. We can take the function easily and flexibly to define the hierarchy of library, branches, offices and divisions.

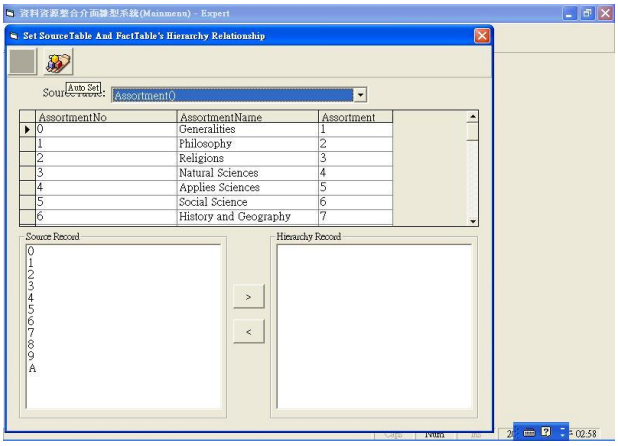

Figure 7. Concept hierarchy defining

4. Bibliomining results can be exported to EXCEL as presentation in visual statistical graphs. We take the powerful and easy-usage advantages of EXCEL graphic capability to help present output patterns.

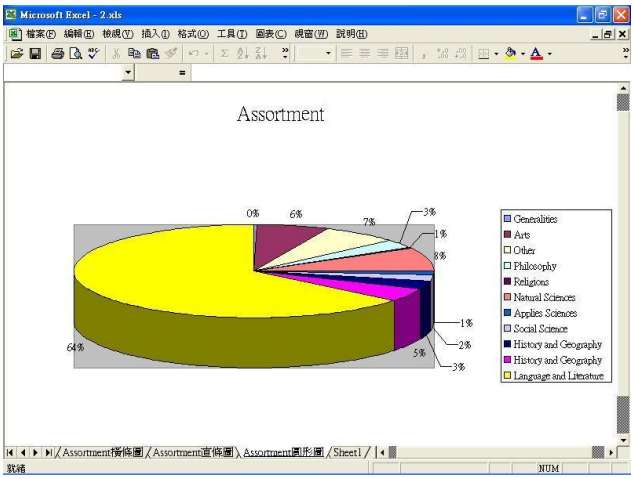

Figure 8. Output Excel visual graph

Conclusions

Bibliomining is a typical garbage-in-garbage-out and highly database technical dependent technique. However, the primary jobs of data mining explorers are to discover what meaningful and useful information to aid their decision makings. They must pay much attention on how to meet their requirements. Thus, librarians and library specialists should have an easy and friendly even free-technology integration system for operating data mining process in order to promise the success of applying bibliomining in libraries. In this paper, we have proposed a novel bibliomining application model and have developed the bibliomining integration prototype system to test and verify the feasibility of proposed model.

References

Atkins, S. (1996). Mining automated systems for collection management. *Library Administration & Management, 10(1)*, 16-19.

Banerjee, K. (1998). Is data mining right for your library? *Computers in Libraries, 18(1*0), 28-31.

Guenther, K. (2000). Applying Data Mining Principles to Library Data Collection. *Computers in Libraries, 20(4)*, 60-63.

Kao, S. C., Hang, H. C. & Lin, C. H. (2003). Decision support for the academic library acquisition budget allocation via circulation database mining. *Information Processing and Management: an International Journal*, *39(1)*, 133-147.

Larsen, P. (1996). Mining Your Automated System for Better Management. *Library Administration & Management*, *10(1)*, 10.

Mancini, D. D. (1996). Mining your automated system for system wide decision making. *Library Administration & Management, 10(1)*, 11-15.

Neumann, A., Geyer-Schulz, A., Hahsler, M., & Thede, A. (2003). An Architecture for Behavior Based Library Recommender Systems. *Information Technology and Libraries*, *22(4)*, 433-454.

Nicholson, S. (2003). The Bibliomining Process: Data Warehousing and Data Mining for Library Decision-Making. *Information Technology and Libraries*, *22(4)*, 146-151.

Nicholson, S. (2006). The basis for bibliomining: frameworks for bringing together usage-based data mining and bibliometrics through data warehousing in digital library services. *Information Processing and Management*, *42*, 785-804.

Peters, T. (1996). Using transaction log analysis for library management information. *Library Administration & Management, 10(1)*, 20-25.

Wu. C.H. (2003). Data mining applied to material acquisition budget allocation for libraries: design and development. *Expert Systems with Applications*, *25(3)*, 401-411.

About the Author

J. C. Shieh earned a Ph. D. in computer science and information engineering from National Taiwan University. He is currently an associate professor of Graduate Institute of Library and Information Studies at National Taiwan Normal University, teaching Information Architecture, Advanced Database Design & Management and Data Warehouse & Data Mining. He is specialized in Bibiliomining, Library Management and Information Architecture on Webs.