The current issue and full text archive of this journal is available at
www.emeraldinsight.com/0264-0473.htm

The integration system for librarians' bibliomining

Jiann-Cherng Shieh
Graduate Institute of Library and Information Studies, National Taiwan Normal University, Taipei, Taiwan

Received 25 May 2009
Revised 29 September 2009
Accepted 5 October 2009

Abstract

Purpose – For library service, bibliomining is concisely defined as the data mining techniques used to extract patterns of behavior-based artifacts from library systems. The bibliomining process includes identifying topics, creating a 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 successful data mining applications. However, the data warehouse creation in the processing of various data sources obviously hampers librarians to apply bibliomining to improve their services and operations. Moreover, most market data mining tools are even more complex for librarians to adopt bibliomining. The purpose of this paper is to propose a practical application model for librarian bibliomining, then develop its corresponding data processing prototype system to guarantee the success of applying data mining in libraries.

Design/methodology/approach – The rapid prototyping software development method was applied to design a prototype bibliomining system. In order to evaluate the effectiveness of the system, there was a comparison experiment of accomplishing an assigned task for 15 librarians.

Findings – With the results of system usability scale (SUS) comparison and turn-around time analysis, it was established that the proposed model and the developed prototype system can really help librarians handle bibliomining applications better.

Originality/value – The proposed novel application bibliomining model and its developed integration system are proved to be effective and efficient in bibliomining by the task-oriented experiment and SUS to 15 librarians. Comparing turn-around time to accomplish the assigned task, about 35 per cent in terms of time was saved. Librarians really require an appropriate integration tool to assist them in successful bibliomining applications.

Keywords Data mining, Libraries, Librarians, Data management

Paper type Research paper

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 (Banerjee, 1998; Nicholson, 2003, 2006). In order to meet the needs of different patron groups, libraries can apply the data mining process to uncover patterns or rules of artifacts of use in communities that gain library services (Nicholson, 2003; Neumann *et al.*, 2003). Furthermore, data mining

The Electronic Library
Vol. 28 No. 5, 2010
pp. 709-721

0264-0473
DOI 10.1108/02640471011081988

can also be applied to discover effective information from library operation-related data sources to aid in the support of library management decision makings (Atkins, 1996; Guenther, 2000; Kao *et al.*, 2003; Larsen, 1996; Peters, 1996; Wu, 2003).

In the process of library 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 obvious to realize that a properly constructed and consolidated data warehouse is the critical promise to success data mining applications. For specified purposes, some current extraction, transformation and loading (ETL) tools help database experts do the job well. However, these highly database techniques and dependent tasks will hamper librarians to take to bibliomining to improve their operations or services flexibly. Librarians have fruitful domain knowledge but they are often somewhat weak in the technical capabilities of doing data extraction, data transformation and data cleaning tasks.

For bibliomining to be easy and smooth for librarians, we really need a user-friendly integration system of gathering data sources and interfacing data mining tools for them. Using the system, we can encourage librarians to apply data mining techniques to provide useful and necessary information for their management requirements, focusing on the professional issues but without interference caused by database techniques. This is highly important and critical to promote the bibliomining applications in libraries.

In order to achieve an easily usable and flexible application of bibliomining for librarians, in this paper we first propose a novel bibliomining application model which is more appropriate for library operations. We then develop its corresponding integration system of bibliomining by a rapid prototyping method. Taking a system usability scale (SUS) and a task-oriented experiment involving 15 librarians on the developed system and a SQL server-based environment, we have shown that the system is effective and efficient for librarians' bibliomining in turn-around time as well as usage satisfaction.

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 shown in Figure 1: 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 mining 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.

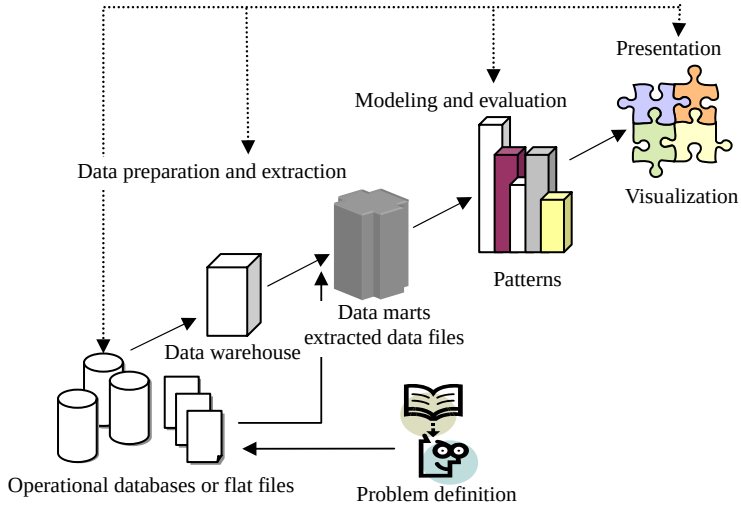

Figure 1.
The process of data mining

Bibliomining

Data mining techniques can help libraries in knowing the trends of popular subjects to enable them to have a 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 in order 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 (Atkins, 1996; Chen and Chen, 2007; Guenther, 2000; Kao *et al.*, 2003; Larsen, 1996; Peters, 1996; Tsai and Chen, 2008; Wu, 2003). 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 anonymizing 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 (Nicholson, 2003).

Bibliomining is just another synonym for library data mining with different experts involved in the process. Domain experts, in this case librarians or library specialists, identify required data sources to provide specific services, to resolve management issues or to help decision makings in the 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 the proper tools to discover meaningful patterns to predict or to describe different patrons or clusters of demographic groups that exhibit certain characteristics. Sometimes, libraries only have data marts or extract data from data resources to do bibliomining for special dedicated issues. The general bibliomining application process is shown in Figure 2.

For example, librarians would like to improve the efficiency of circulation services by considering different workloads. They should know how to arrange manpower on the desk. Library specialists and librarians may have ideas to analysis the time spans

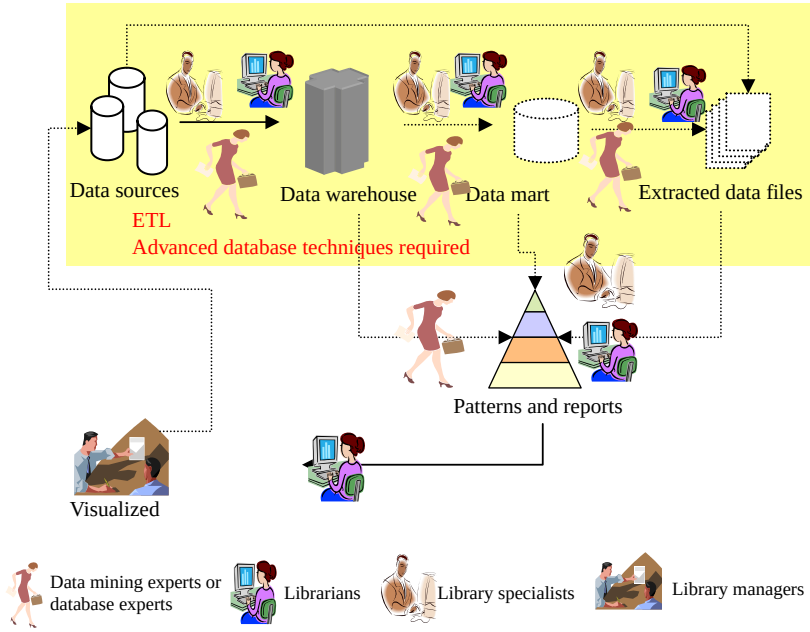

Figure 2.
The application flow of bibliomining

of historical circulations to get helpful information. They thus identify borrowing records and patrons' data for the previous five academic years as required data resources 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, for example, library automation systems.

However, the formats or types of fields used to record circulation data are not quite equal completely, such as ten characters for K's student ID but eight for S's. Data mining experts or database specialists must do data conversion by SQL programming to resolve the issue. After the 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 and obtain the results. The mining process will be iterated until the resulted information is verified and proved by library specialists and librarians, and library managers. Only then is the bibliomining case finished.

The above bibliomining result is providing information to improve the circulation services only. If now we have another service issue about book recommendations for patrons, we must go through the previous steps again with integrating different required data. Librarians are weighed down with the bibliomining work and are completely exhausted. The bibliomining studies were all case-by-case applications of collections, budget allocation, acquisitions, books recommendations, etc. (Atkins, 1996; Chen and Chen, 2007; Kao *et al.*, 2003; Tsai and Chen, 2008; Wu, 2003).

In the mentioned scenario, data mining or database experts play significant roles in the process, integrating various data resources, transforming data, extracting data, creating data warehouse and selecting mining algorithms. These technical tasks require strong database disciplines, but librarians generally have poor knowledge of these.

Librarians are thus frustrated by repeated failures as attempting to apply bibliomining. Furthermore, these case-by-case applications will make everyone annoyed and impatient.

In practical implementation, 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 or disclose interesting patrons' behaviors. Even though many diverse ETL tools exist, the database specialists are ham-fisted to take data collection, ETL to build initial data warehouses required by data mining case by case. The situation really obstructs the flexibility of bibliomining adoption in libraries.

In the next section, we will propose a novel bibliomining application model which is more appropriate for librarians. Based on this model, we will develop its corresponding integration prototype system. The system is centered on librarians' requirements in applying bibliomining to improve their services. The system provides user-friendly interfaces that are transparent to librarians with less database programming literacy. We expect that the system will facilitate bibliomining applications in libraries.

Bibliomining application model

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

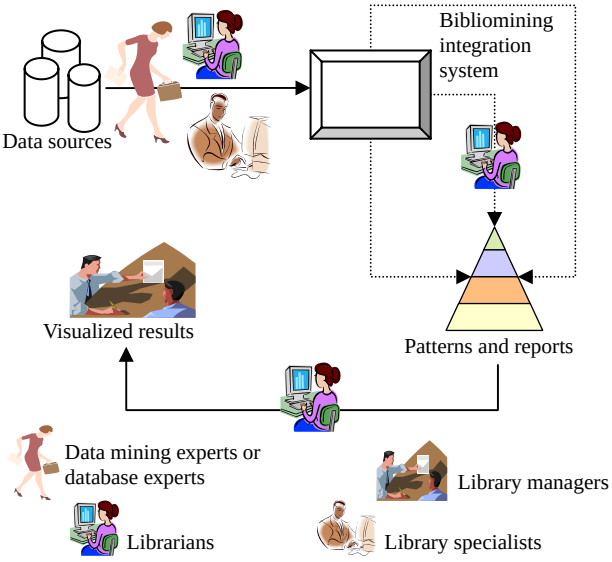

Figure 3.
Bibliomining application model

For integrating all possible data sources, data mining experts and library specialists can define relationships among them comprehensively at first. Data sources formats can be databases, Excel files and text files and together they create an overall data warehouse for a specific library. When librarians propose a designated problem for bibliomining, they keep their minds focused on understanding and resolving the issues through their library domain knowledge. After confirming the relating data required, librarians use the bibliomining integration system to do the drilling and picking operations on the needed data items or fields to constitute the problem-oriented data cube. Self-defining generalized concept hierarchies, librarians can classify, categorize, cluster data as they require for further analysis purposes. Librarians finally generate and present the results. It is clear that the bibliomining integration system facilitates the librarians' operations without worrying them about the absence of database technology.

Next, we should develop the bibliomining integration system to support the model. The bibliomining integration system has two main types of functions. For experts and specialists, the system provides functions to integrate data sources, to identify data relationships, to diagram data warehouse models (star schema and snowflake schemata), to transform and load data, to create comprehensive data warehouses and so on. For librarians, the system supplies functions to define data concept hierarchy, to select required data items, to create specific cube, to import data to online analytical processing or data mining tools and to output the visual results or patterns. Data mining experts and library specialists will have much responsibility of integrating critical and necessary data sources to construct an overall data warehouse as possible. 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 shown in Figure 4.

In summary, we can say:

- The system would handle various data resources such as different databases or data files for library bibliomining. Based on domain knowledge experiences, library specialists confirm all possible required data resources for a specific library's applications. Data mining or database experts then do data integration from specific database servers and tables that required data resides in.
- With the help of database experts, the non-database data can be transformed and imported from text files to constitute tables. In the meantime, database experts and library specialists can define new fields for required data that is never considered to collect at previous design.
- According to data correlations, library specialists set relationships for some new created tables or fields. We have prepared all possible bibliomining required data tables for the designated library.
- Database experts and data mining experts can build a comprehensive and overall data warehouse for the designated library. Whenever librarians have defined their interesting issues, they can select specific data from the data warehouse to create data marts or cubes for solving the purposed problems. Particularly, they can self-define concept hierarchies of selected data of digit, date and characters types for further analysis need.
- After creating the problem-oriented cubes, librarians can export them to data mining or OLAP tools to generate text reports or visual graphics results.

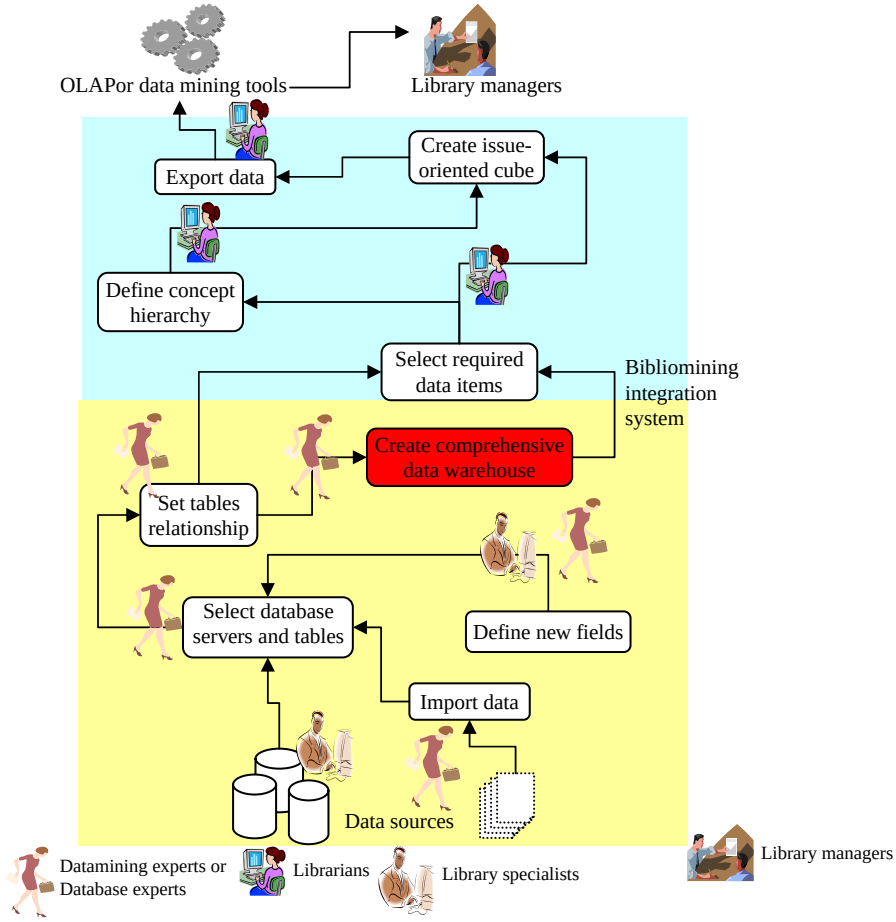

Figure 4.
Bibliomining integration system structure

The bibliomining integration system

Based on the proposed bibliomining application model and system structure, we develop the bibliomining integration prototype system using rapid prototyping development methods. The prototype is developed by MS Visual Basic and SQL Server 2005 at Windows Server 2003. The system can handle various data sources, such as MS Access, SQL Server, MS Excel and text files. Especially, self-defining concept hierarchy function provides librarians to clarify aggregated data in different slices as their analysis requirements. Concept hierarchies are constructed from selected data of character, date and number types. For example, based on names of academic divisions running at a university, we can define a hierarchy of schools, departments and divisions. The system supports star schemata and snowflake schemata to build data warehouses and data cubes. Librarians can build the schemata by click and drag. After completing the required data warehouse schemata, we can make system export the corresponding data to data mining or OLAP tools, such as MS Excel, SQL Server or SPSS Clementine for further processing.

For integrating data sources defined by library specialists, the system provides the interface for selecting required tables from various databases resided in different database management systems of the specific servers. In the snapshot (Figure 5) "Student" database is selected as the secondary database for defining dimensional tables and then necessary data fields can be picked up. We can also define the main database where fact tables are located in through the snapshot.

Based on the selected and defined data tables, data items, fact tables, dimension tables and tables' relationships, the integration system generates the snowflake database schema as snapshot shown in Figure 6. A noted, currently, the system provides both snowflake and star schemata creation.

What critical function the system provides is dependent on the concept hierarchies users require. The system can help users define these concept hierarchies from table fields associated with various types of data such as date, number or character. Figure 7 gives the snapshot of concept hierarchy creation from the table field of character data.

Finally, the system generates the results with appropriate data formats to feed data mining, OLAP or visualization tools for further processing. Figure 8 shows the snapshot of exporting the results to Excel for presentation in a visual statistical graph. We take the powerful and easy-usage advantages of the Excel graphic capability to help present output patterns.

The experimental analysis

In order to verify the developed prototype's effects, we take a task-oriented experiment in which 15 librarians have individually tried to accomplish the assigned task of creating data marts on the prototype integration system and SQL server analysis services, respectively. Computing turn-around time to complete the task and filling out questionnaires of SUSs,

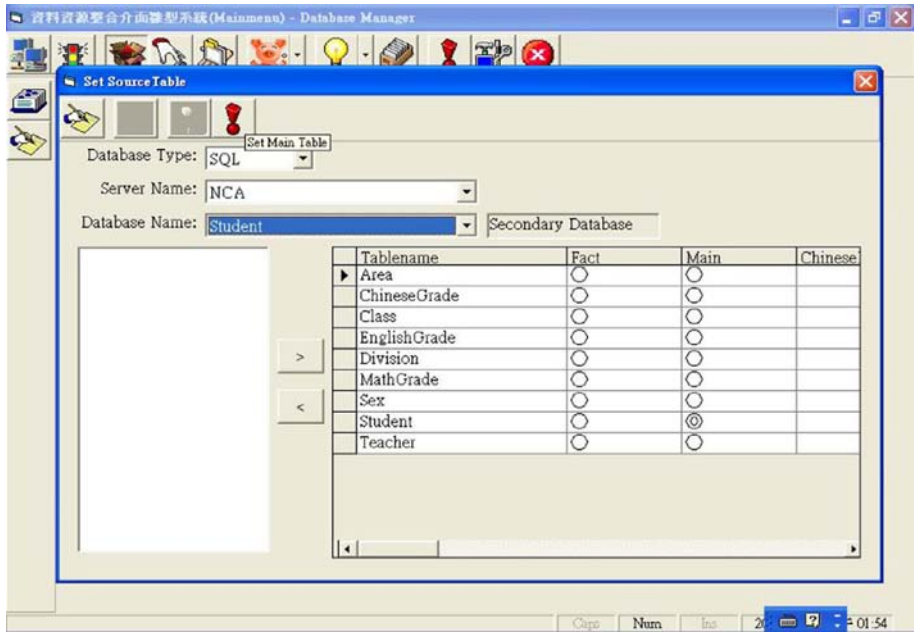

Figure 5.
Databases and tables selection

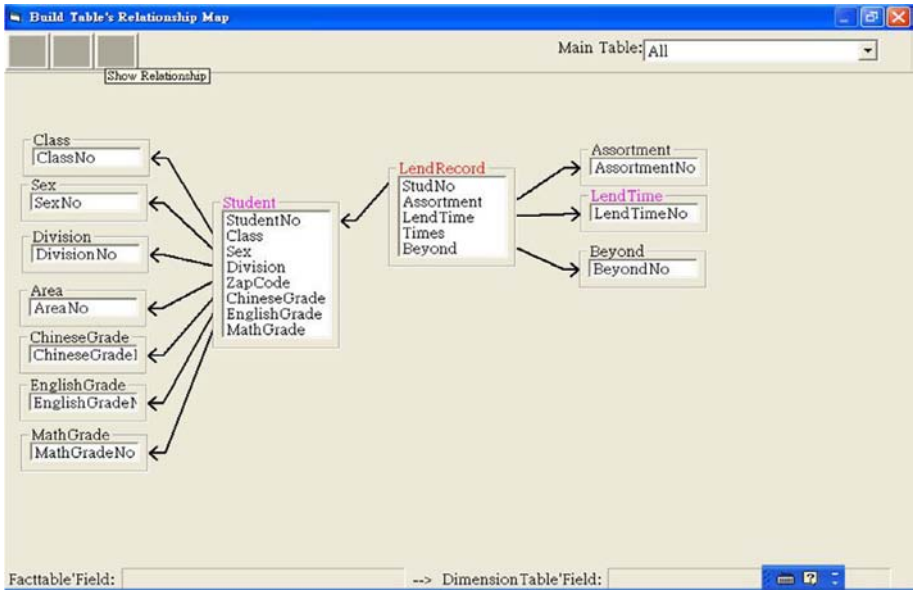

Figure 6.
Snowflake schema model building

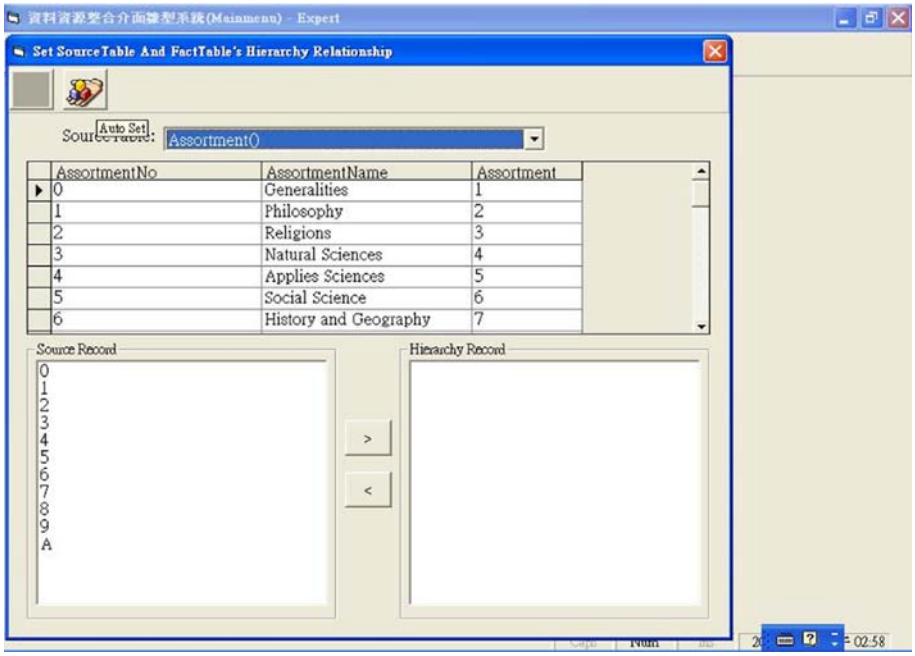

Figure 7.
Concept hierarchy creation

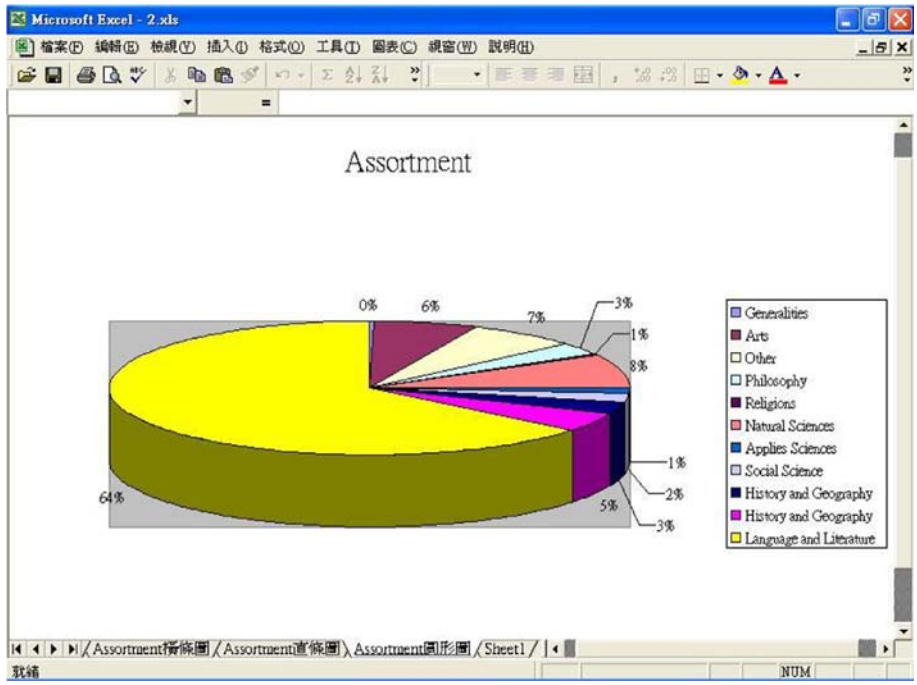

Figure 8.
Output Excel graphic functions for information visualization

we have their SUSs comparison and turn-around time analysis on them. SUS is used to verify the usability of the developed prototype system and SQL Server Analysis Services. The task-oriented experiment is to measure their effectiveness and efficiency in terms of turn-around time to accomplish the assigned task. By the analysis of variance (ANOVA), we expect to prove that they are statistically significant differences in librarians' bibliomining applications.

System usability comparison

We apply SUS (Brooke, 1996) to measure the comparison of the usability of the developed prototype system and SQL Server Analysis Services. SUS is a simple, ten-item Likert scale giving a global view of subjective assessments of usability. SUS yields a single number representing a composite measure of the overall usability of the system being studied. Note that scores for individual items are not meaningful on their own. To calculate the SUS scores, first sum the score contributions from each item. For items 1, 3, 5, 7 and 9 the score contribution is the scale position minus 1. For items 2, 4, 6, 8 and 10, the contribution is 5 minus the scale position. Multiply the sum of the scores by 2.5 to obtain the overall value of SUS. Table I shows the SUS scores of the 15 librarians.

We apply Excel 2003 data analysis ANOVA: single factor function to analyze the SUS experimental scores data. The results are shown in Table II and it is plain to see that the p-value ($= 0.003576757$) is much < 0.05. This concludes that IS and AS have statistically significant differences. IS achieves better average performance at system usability.

Task-oriented test

During the experiment, we also recorded the time that each librarian takes to create data marts to complete the task. Table III shows the turn-around time data. In this task case, we have taken 35 per cent ((151.14 − 97.57)/151.14 = 0.354) less time to accomplish the task in average. Next, we analyzed the time data by Excel 2003 Data Analysis ANOVA: single factor function. Table IV shows the analytical results. It can be concluded that IS and AS are significantly different on turn-around time issue since *p*-value (= 0.042651962) is <0.05. And IS has less average turn-around time.

Librarians	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
IS[b]	65	70	45	67.5	72.5	70	57.5	80	95	70	NA[a]	65	72.5	62.5	77.5
AS[c]	35	65	42.5	62.5	55	57.5	62.5	57.5	57.5	52.5	NA[a]	45	75	62.5	57.5

Notes: [a]Number 11 librarian gave up the test during the experiment; [b]IS represents "The prototype Integration System"; [c]AS represents "MS SQL Server Analysis Services"

Table I. Table1 SUS scores

Summary						
Groups	*Count*	*Sum*	*Average*	*Variance*		
IS	14	970	69.28571429	129.2582418		
AS	14	787.5	56.25	102.6442308		
ANOVA						
Source of variation	*SS*	*df*	*MS*	F	p-*value*	F *crit*
Between groups	1,189.509	1	1,189.508929	10.25869984	0.003576757	4.22520119
Within groups	3,014.732	26	115.9512363			
Total	4,204.241	27				

Table II. SUS scores ANOVA analysis

Librarians	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
IS[b]	49	162	118	103	40	23	187	128	31	44	NA[a]	135	187	99	60
AS[c]	282	251	131	109	251	90	175	158	101	127	NA[a]	195	155	54	29

Notes: [a]Number 11 librarian gave up the test during the experiment; [b]IS represents "The prototype Integration System"; [c]AS represents "MS SQL Server Analysis Services"

Table III. Turn-around time in minutes

Summary						
Groups	*Count*	*Sum*	*Average*	*Variance*		
IS	14	1,366	97.57142857	32,97.648352		
AS	14	2,116	151.1428571	5,544.593407		
ANOVA						
Source of variation	*SS*	*df*	*MS*	F	p-*value*	F *crit*
Between groups	20,089	1	20,089.28571	4.543934958	0.042651962	4.22520119
Within groups	114,949	26	4,421.120879			
Total	135,038	27				

Table IV. Turn-around time ANOVA analysis

From the above experimental results and analysis, we have evidence to conclude that the proposed bibliomining model and the developed prototype system can really help librarians do bibliomining well and efficiently.

Conclusions

Bibliomining is certainly defined to be a highly information and computer-dependent technique applied in libraries. The primary and much important job of bibliomining is to discover what meaningful and useful information patterns can help library managers in decision making. Bibliomining significantly concerns professional knowledge. Attention must be paid on how to uncover critical information and to meet their requirements closely, and to identify what necessary data they require to solve specific problems. However, the threshold of required advanced database techniques to process initial data may become serious barriers for librarians' bibliomining applications. In this paper, we have proposed a novel bibliomining application model to help librarians overcome any embarrassing and unassisted situation. Adopting a rapid prototyping development method, we have built the corresponding integration system to make bibliomining in libraries possible and flexible. We have undertaken SUSs and a task-oriented experiment to prove the proposed novel mode and developed prototype system workable and effective. The time saved in the experiment was as much as 35 per cent to accomplish the assigned task in average and get better performance in system usability. This alone shows the efficacy of the proposed system.

References

Atkins, S. (1996), "Mining automated systems for collection management", *Library Administration & Management*, Vol. 10 No. 1, pp. 16-19.

Banerjee, K. (1998), "Is data mining right for your library?", *Computers in Libraries*, Vol. 18 No. 10, pp. 28-31.

Brooke, J. (1996), "SUS: a 'quick and dirty' usability scale", in Jordan, P.W., Thomas, B., Weerdmeester, B.A. and McClelland, A.L. (Eds), *Usability Evaluation in Industry*, Taylor & Francis, London.

Chen, C.C. and Chen, A.P. (2007), "Using data mining technology to provide a recommendation service in the digital library", *The Electronic Library*, Vol. 25 No. 6, pp. 711-24.

Guenther, K. (2000), "Applying data mining principles to library data collection", *Computers in Libraries*, Vol. 20 No. 4, pp. 60-3.

Kao, S.C., Hang, H.C. and Lin, C.H. (2003), "Decision support for the academic library acquisition budget allocation via circulation database mining", *Information Processing & Management: an International Journal*, Vol. 39 No. 1, pp. 133-47.

Larsen, P. (1996), "Mining your automated system for better management", *Library Administration & Management*, Vol. 10 No. 1, p. 10.

Neumann, A., Geyer-Schulz, A., Hahsler, M. and Thede, A. (2003), "An architecture for behavior based library recommender systems", *Information Technology and Libraries*, Vol. 22 No. 4, pp. 433-54.

Nicholson, S. (2003), "The bibliomining process: data warehousing and data mining for library decision-making", *Information Technology and Libraries*, Vol. 22 No. 4, pp. 146-51.

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 & Management*, Vol. 42, pp. 785-804.

Peters, T. (1996), "Using transaction log analysis for library management information", *Library Administration & Management*, Vol. 10 No. 1, pp. 20-5.

Tsai, C.S. and Chen, M.Y. (2008), "Using adaptive resonance theory and data-mining techniques for materials recommendation based on the e-library environment", *The Electronic Library*, Vol. 26 No. 3, pp. 287-302.

Wu, C.H. (2003), "Data mining applied to material acquisition budget allocation for libraries: design and development", *Expert Systems with Applications*, Vol. 25 No. 3, pp. 401-11.

Further reading

Mancini, D.D. (1996), "Mining your automated system for system wide decision making", *Library Administration & Management*, Vol. 10 No. 1, pp. 11-15.

About the author

Jiann-Cherng Shieh earned a PhD in computer science and information engineering from the National Taiwan University. He held positions as library directors of Nanhua University and Fo-Guang University in Taiwan. He is currently an Associate Professor at the Graduate Institute of Library and Information Studies at National Taiwan Normal University where he teaches information architecture, advanced database design and management and data warehousing and data mining. Besides these topics he also specializes in bibiliomining and information ethics. Jiann-Cherng Shieh can be contacted at: jcshieh@ntnu.edu.tw