Association Rules Mining for Knowledge Management : A Case Study of Library Services

Chu Chai Henry Chan* Ming-Hsiu Lee Yun-Chiang Kwang

E-Business Research Lab
Department of Industrial Engineering and Management,
Chaoyang University of Technology
No.168, Jifong East Road, Wufong Township, Taichung County, Taiwan, R.O.C.

Abstract: - Data mining has been applied successfully in a lot of business communities for understanding and tracking behavior of individual or certain groups. To realize the actual needs of college students, this study proposes using data mining to discover the association rules of a library database. This major advantage of the study is to provide a novel mechanism by using problem-solving oriented approach rather than technical concept done by most of previous researches. We apply the Apriori algorithm as the core methodology of implementing association rules mining. To prove the proposed methodology, an empirical case study is conducted to find the association between different users' demands. Moreover, for knowing about students' preference, this work finds the association rules and searches the top ten ranking of books for students of three different colleges. One interesting finding is that different college students have different needs and behavior patterns. This conclusion can give a guideline for the studied library to understand the needs of different background students. Following the finding, the studied university can offer students suitable services in the future.

Key-Words: - Data Mining, Association Rule, Apriori Algorithm, Knowledge Management

1 Introduction

Currently, libraries in most universities are providing a lot of online services [23]. Offering the right service to the right user at the right time is the key which can manage digital library efficiently. Even data mining have applied in many business areas [2, 24, 29]. Unfortunately, this technology still rarely has used in academic library services [17]. Libraries are the major sources of knowledge for all the schools. The efficiency of libraries will affect the quality of delivering knowledge to the young generation. To assist managing library efficiently, this study proposes using data mining to discovering the associate rules of library users' behavior [8, 25, 28]. Furthermore, this study defines a behavior model of library knowledge [1, 18, 19]. Based on the behavior model, this investigation identifies the associate rules of a library database. Following the found associate rules, the studied library can understand the relationships between different background students and then offer right services in the future. To conduct the task of data mining, this study uses the Apriori algorithm to find out the associate rules from a database containing the information of more than fifty thousand library users [8]. The most interesting part of this study is to observe the patterns of library users. We are very curious about two major issues. First, Does different background student have different demand from library? If their demands are different, what are the association rules when students borrow books? Second, what are the most popular books required by students? Can we offer right books based on real statistical number instead of experiences?

2 Data mining

Empirically, many companies face a painful problem that loses their most profitable users daily. Many previous researches began to use data mining as a tool to find out the hidden problems of losing user satisfaction [6, 9, 21]. The definition of data mining is summarized in table 1[3, 13, 16, 24, 25]. The technology of data mining is defined as a sophisticated data search capability that uses statistical or intelligent algorithms to discover trend, patterns and relationship in database [16]. The applications of data mining include discovery clusters, discovering associations, discovering sequential patterns, predicting a classification, predicting customer values and similar time sequences [12, 14].

The popular methods for data mining are a neural network, tree induction, Aprioir algorithm and so on.

Table 1 The Definition of Data Mining

Author	Definition
Sung HO Ha, Sang Chan Park(1998)	The term data mining has different uses in academia and in the commercial marketplace. Data mining is a non-trivial process of identifying valid, novel , potentially useful and ultimately understandable patterns in data.
T.P. Hong, C.S. Kuo, and S.C. Chi (1999)	The goal of data-mining is to discover important associations among items such that the presence of some items in a transaction will imply the presence of some other items.
Indranil Bose, Radha K.Mahapatra (2001)	Data mining is a discipline of growing interest and importance, and an application area that can provide significant competitive advantage to an organization by exploiting the potential of large data warehouses.
Chris Clifton, Bhavani Thuraisingham(2001)	For data mining to be effective, several technologies have to work together. First of all, statistical analysis and machine-learning techniques have to be applied successfully to databases to extract patterns and to predict trends.
Jeffrey W. Seifert(2004)	Data mining involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data sets.

3 User Behavior Model for Decision Making of Library Knowledge

The goal of this study is to generate association rules of a library database. Before discovering association rules of behavior patterns, we have to know the basic definition of a customer behavior model. According to the model of EC (Electronic Commerce) consumer behavior defined by Turban, the decision making is basically a customer's reaction to stimuli [7].

The decision making process is influenced by the characteristics of users, the environment, stimuli and the delivering system. In a library system, librarians often neglect the real need of users. To cope with such a problem, this study chooses several important factors of stimuli (course or research need, personal interest) and personal characteristics (education background) to describe the behaviors of customers..

4 Data Mining Processes

To complete the task of data mining, this study develops five-stage mechanism (see Fig.1) to find the association rules of library knowledge [10, 20, 26, 27]. First, data has to be collected into a database [1, 4, 19]. Second, data has to be normalized and processed as the format which can be analyzed [3]. Third, data warehousing technique is used to convert data which can be viewed by a multi-dimensional database [5]. Four, a data-mining tool are used to analyze data to discovery association rules or decision trees [22]. Finally, useful knowledge or intelligence is summarized and represented to users.

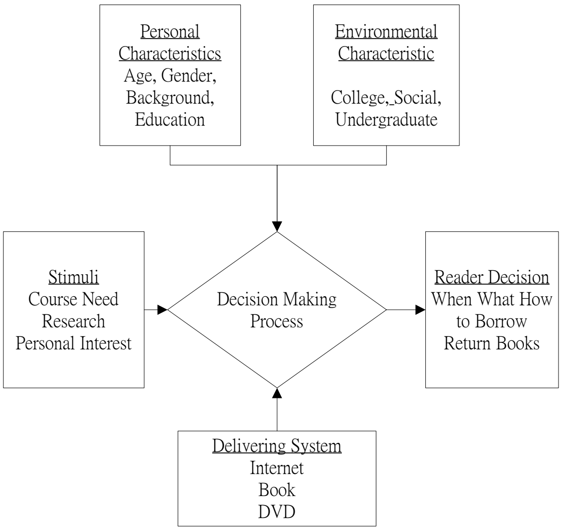

Fig. 1 Behavior Model of Library Users

5 Apriori Algorithm

This work uses the Apriori algorithm as a methodology to find the association rules of a library database. Just like a lot of previous researches mention, the original problems of association rules come from supermarkets which analyze the associated relationships between two different items purchased in a basket [28, 8]. Based on the results of analysis, managers can arrange a new layout or promotion plan to increase the value of users [6]. Three major terms for the Aprioir algorithm are Support, Confidence and Improvement. The basic definitions of Support, Confidence and Improvement are as follows [8]:

The support of rul**e X => Y is** the percentage of transaction which contain both X and Y.

Support **(**X =>Y**)** =

$$\frac{\text{(\# of transaction containing both X and Y)}}{\text{((total \# of transaction containing in the database)}} \quad (1)$$

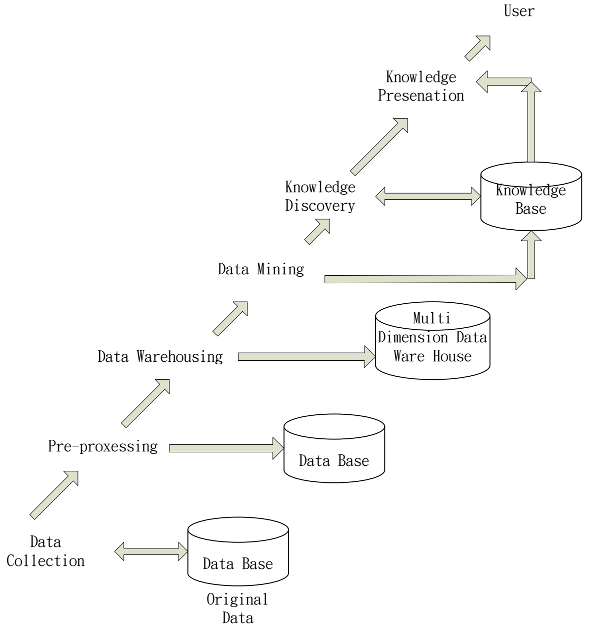

Fig. 2 Data Mining Processes for Knowledge Discovery

The Confidence of rule X => Y is defined as the measure of certainty or trustworthiness associated with transaction X [14, 30].

Confidence **(**X =>Y**)** =

$$\frac{\text{(\# of transaction containing both X and Y)}}{\text{(\# of transaction containing X)}} \quad (2)$$

The Improvement of rule X => Y is defined as the measure of the possibility of X and Y both happening.
Improvement (X =>Y) =

$$P（X\cap Y）/P（X）*P（Y） \quad (3)$$

The process of mining association rules is a three-stage approach which determines the large itemset of association rules by support, confidence and Improvement(See Fig. 2). The purpose of Association Rule Mining is to find all frequent itemsets and generate strong association rules from the frequent itemsets. The concept of Apriori Algorithm includes following functions [14, 30]:

- Apriori algorithm is an influential algorithm for mining frequent itemsets for Boolean association rules.
- Uses a Level-wise search, where k-itemsets (An itemset that contains k items is a k-itemset) are used to explore (k+1)-itemsets, to mine frequent itemsets from transactional database for Boolean association rules.
- •Apriori algorithm is an influential algorithm for mining frequent itemsets for Boolean association rules.
- First, the set of frequent 1-itemsets is found. This set is denoted L1. L1 is used to find L2, the set of frequent 2-itemsets, which is used to fine L3, and so on, until no more frequent k-itemsets can be found Association rule mining process
- Find all frequent itemsets:
- Each support **S** of these frequent itemsets will at least equal to a pre-determined min_sup (An *itemset* is a subset of items in I, like A)
- Generate strong association rules from the frequent itemsets:
- These rules must be the frequent itemsets and must satisfy min_sup and min_conf.

6 Empirical Study

A lot of libraries like to follow the rules decided by top managements instead of responding the actual needs of students. To solve this problem, this study uses data mining approach to analyze student's behavior for discovering their potential needs. The proposed approach shown in Fig. 3 includes several stages: problem acquisition, data collection, problem definition, data preparation, model & hypothesis development, data mining, knowledge discovery and problem solving. In the first stage, we have to acquire key problems happening currently. To verify the problems, problem definitions have to be made and related data have to be prepared. The third step is to make hypothesis for problems and develop an associated model. The fourth step is to apply data mining to generate useful knowledge for solving problems. Finally we use valuable intelligences or knowledge to users for solving their problems. By following the mechanism, we acquire the problems of this study as follows:

- How can a library use limited budget to satisfy the actual needs of students?
- How can libraries assist students to find their required books as soon as possible?

According to two above problems, the expectations for this investigation are summarized as two following missions:

- Know about the ranking of students' preference
- Understand the relationship between different records in a library

All the details of information are shown in table 2. This study collects 58959 records held between January 2006 and December 2006 from a university library. We use Poly-analysis 4.6 software to find the association rules between transaction data. In this case study, we define minimum Support, minimum Confidence and minimum Improvement as 1.00%, 30.00% and 2.00%. It means that this study will choose the rules with the values of Support, Confidence and Improvement of any rule bigger than 1.00%, 30.00% and 2.00%.

Table 3 shows the associations rules of students of the social science college. For example, if a student of the social science college borrows the books of Poetry or Chinese literature or Biography of China, then he/she may borrow the books of Chinese rhyme with support =2.64%, Confidence=65.96% and Improvement= 8.996. Similarly, table 4 shows the association rules of students of the management college. The highly association rule is Poetry=> Rhyme. It means if one student of the management college borrows the books of Poetry, then it is highly potential that he/she will borrow the books of Rhyme. Table 5 shows the association rules of the engineering college. The highly association rule is Organic chemistry => Analytical chemistry. It indicated if one student of engineering college borrows the books of Organic chemistry, then he/she will borrow the books of analytical chemistry. Based on the experimental result, one interesting finding is that different college students have different need and behavior patterns. This study strongly suggestions the studied library should put the highly associated books together. This way can assist students to find their required book rapidly.

To know about the ranking of students' preference, this study searches the top ten ranking of books for students of three different colleges. If the budget of libraries is tight, they can buy those high-ranking books such as math, computer, fiction and literature. This outcome will provide a simple guide to the studied library to purchase right books for students.

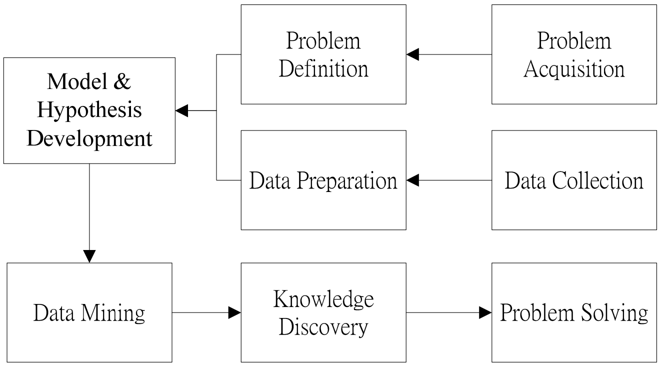

Fig. 3 Data Mining Processes for Problem Solving

Table 2 Description of Problem Solving by Data Mining

Problem Solving Stages	Details
Problem Statement	● How can an university library use limited budget to satisfy the actual needs of students? ● How can a library assist students to find their required books as soon as possible?
Expectation	● Know about the ranking of students' preference ● Understand the relationship between different records in a library
Data Source	58959 records in a library database held between January 2006 and December 2006
Tool	Poly-analysis 4.6
Model	Data Mining to Discover Association Rules & Decision Tree

Table 3 Association Rules of Social Science College

Association Rules		Support	Confidence	Improvement
		> 1.00%	> 30.00%	> 2.00%
Poetry, Chinese literature, Biography of China	-> Chinese rhyme	2.64%	65.96%	8.996
Chinese rhyme, Chinese literature, Biography of China	-> Poetry	2.64%	62.00%	9.091
Chinese rhyme, Poetry, Biography of China	-> Chinese literature	2.64%	63.27%	6.289
Chinese rhyme, Poetry, Chinese literature	-> Biography of China	2.64%	64.58%	3.443
Social control, Education, Family	->Social work	1.11%	36.11%	6.322
Social work, Education, Family	-> Social control	1.11%	48.15%	5.945
Social work,	-> Education	1.11%	39.39%	11.552

Association Rules		Support	Confidence	Improvement
		> 1.00%	> 30.00%	> 2.00%
Social control, Family				
Social work, Social control, Education	-> Family	1.11%	54.17%	2.875
Educational psychology, General education	-> Education	1.96%	38.33%	2.661
Educational psychology, Education	-> General education	1.96%	31.51%	5.280
General history of china	-> Han and Three kingdoms history	3.24%	35.19%	6.069

Table 4 Association Rules of Management College

Association Rules		Support	Confidence	Improvement
		> 1.00%	> 30.00%	> 2.00%
Poetry	->Rhyme	1.27%	60.00%	40.457
Rhyme	-> Poetry	1.27%	85.71%	40.457
American literature	->English literature	6.78%	42.67%	3.729
English literature	-> American literature	6.78%	59.26%	3.729
Insurance	->Finance	5.51%	40.00%	3.255
Finance	->Insurance	5.51%	44.83%	3.255
Marketing	->Business management	2.30%	65.85%	6.280

Table 5 Association Rules of Engineering College

Association Rules		Support	Confidence	Improvement
		> 1.00%	> 30.00%	> 2.00%
Applied psychology	-> Psychology of personality	3.14%	44.44%	13.060
Psychology of personality	-> Applied psychology	3.14%	92.31%	13.060
Organic chemistry	-> Analytical chemistry	2.36%	60.00%	12.733
Analytical chemistry	-> Organic chemistry	2.36%	50.00%	12.733
Japan literature	-> American literature	5.50%	58.33%	4.369
American literature	-> Japan literature	5.50%	41.18%	4.369
Physical chemistry	->Inorganic chemistry	1.57%	37.50%	23.875
Inorganic chemistry	-> Physical chemistry	1.57%	100.00%	23.875

7 Discuss and Conclusion

Even the theory of data mining has been developed and used in business sectors for many years. Not many works has done for schools to assist analyzing the behavior of students or faculties. Besides, most of previous researches related data mining focuses on theoretical study. This study attempts to use data mining to solve the real problems of users in a library. The advantages of the proposed approach are

- Problem-solving oriented rather than conventional techniques oriented
- Data Mining can be used as a practical technique instead of a profound theory

In this study, we utilize the Apriori algorithm to mine the association rules of a library database. We collect 58959 records held between January 2006 and December 2006 from an university library and use Poly-analysis 4.6 software to find the association rules of transaction data. One interesting finding is that different background students have different need and behavior. For instant, if a social science student borrows a book related to a book of Poetry, then he/she will borrow a book of Chinese rhyme later. In the same situation, a student of the management college will borrow a book of Rhyme instead of Chinese rhyme. The conclusion can give a guideline for libraries to understand the need of different background students. The studied university can offer different students suitable services in the future.

8 Acknowledgments

The authors would like to thank the National Science Council of the Republic of China, Taiwan for financially supporting this research under Contract No. NSC95-2221-E-324-032

References:

[1] B. Mobasher, H. Dai, T. Luo, and M. Nakagawa, “Discovery and Evaluation of Aggregate Usage Profiles for Web Personalization”, Data Mining and Knowledge Discovery, Vol. 6, 2002, pp.61-82.

[2] Chris Rygielski , Jyun-Cheng Wang and David C. Yen,”Data mining techniques for customer relationship management”, Technology In Society,Vol.24, 2002, pp.483-502.

[3] Chris Clifton and Bhavani Thuraisingham, ‘Emerging standards for data mining,’ Computer Standards & Interface, 23, 2001, pp.187-193.

[4] D. J. Hand, G. Blunt, M. G. Kelly and N. M. Adams, ”Data mining for fun and profit”, Statistical Science, Vol.15, No.2, 2000, pp.111-131.

[5] E. Rafalski, “Using Data Dining/Data Repository Methods to Identify Marketing

Opportunities in Health Care," Journal of Consumer Marketing, Vol. 19, 2002, pp. 607-613.

[6] Euiho Suh, Seungjae Lim, Hyunseok Hwang, Suyeon Kim,"A prediction model for the purchase probability of anonymous customers to support real time web marketing: a case study", Expert Systems with Applications,Vol.27, 2004, pp.245-255.

[7] E. Turban, D. King, J.K. Lee and D. Viehland, Electronic Commerce: A Managerial Perspective, Prentice Hall, New Jersey, 2006.

[8] H.C. Lee, Y.H. Kim and P.K.Rhee, "Web Personalization Expert with Combining collaborative filtering and Association Rules Technology," Expert System with Applications, 21, 2001, pp.131-137.

[9] H.M. Chuang and P., Gray, "Special Section: Data Mining", Journal of Management Information Systems, Vol.45, 1999, pp.295-302.

[10] Fernando Crespoa and Richard Weberb," A methodology for dynamic data mining based on fuzzy clustering", Fuzzy sets and systems, Vol. 150, 2005, pp.267-284.

[11] F.H. Grupe. And M.M. Owrang, "Database Mining Discovering New Knowledge and Cooperative Advantage", Information System Management, Vol.12, 1995, pp.26-31.

[12] Frank Chou, "Business Intelligence" IBM, 2001.

[13] Indranil Bose and Radha K. Mahapatra, "Business data mining- a machining leaning perspective," Information & Management, 39, 2001, pp.211-225.

[14] J. Han and M. Kamber, "Data Mining Concepts and Techniques", Morgan Kaufman Publishes AN Imprint of Academic Press, pp.225-354, 2001.

[15] J.B. Schafer, J.A. Konstan, and J. Riedl, "E-Commerce Recommendation Applications," Data Mining and Knowledge Discovery, Vol. 5, 2001, pp.11-32.

[16] Jeffrey W. Seifert, "Data Mining and the search for security Challenges for connecting the dots and databases," Government Information Quarterly,21, 2004, pp.461-480 .

[17] Kyle, B., "Is Data Mining Right for Your Library?," Computers in Libraries, Vol. 18, 1998, pp. 28-31.

[18] Michael E. Welge ," Knowledge management and data mining for marketing", Decision Support Systems, Vol.31, pp.127-137.

[19] M. J. A. Berry and G. Linoff, "Data Mining Techniques: For Marketing, Sales, and Customer Support", John Wiley & Sons, 1997.

[20] M. S. Chen, J. Han and P. S. Yu, "Data Mining: An Overview form a Database Perspective, e," IEEE Trans. On Knowledge and Data Engineering, Vol.8, 1996, pp.866-883.

[21] Olmeda and P. J. Sheldon, "Data Mining Techniques and Applications for Tourism Internet Marketing," Journal of Travel and Tourism Marketing, Vol. 11, 2001.

[22] R. Agrwal, T. a Imielinski and A. Swami, "Database Mining: a Performance Perspective", IEEE Tran. Knowledge and Data Engineering, Vol.5, 1994, pp. 914-925.

[23] Scott Nicholson,"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 ,Vol. 42, 2006, pp.785-804.

[24] Sung HO Ha and Sang Chan Park," Applications of data mining tools to hotel data mart on the intranet for database marketing," Expert System with Applications, 5, 1998, pp.1-13.

[25] T.P. Hong, C.S. Kuo, and S.C. Chi, "Mining Association Rules from quantitative data," Intelligence data analysis, 3, 1999, pp.363-576.

[26] U.M. Fayyad, G. Piatetsky-Shapiro and P. Smithy, "The KDD process for extracting useful knowledge from volumes of data", Communication of the ACM, Vol.39, 1996, pp.27-34.

[27] W. Frawley, G. Piatetsky-Shapiro and C. Maheus, "Knowledge Discovery in database: an overview", AI Magazine, 1992, pp.213-228.

[28] W.Y. Lin, S.A. Alvarez, and C. Ruiz, "Efficient Adaptive-Support Association Rule Mining for Recommender Systems", Data Mining and Knowledge Discovery, Vol.6, 2002, pp.83-105.

[29] W.P. Lee, C.H. Liu, and C.C. Lu, "Intelligent agent-based systems for personalized recommendations in Internet commerce", Expert Systems with Applications, Vol.22, 2002, pp.275-284.

[30] Andrew Kusiak, Association Rules-The Apriori Algorithm, Handout of Intelligent system lab, The University of IOWA, U.S.A.