An Empirical Study of Application of Data Mining Techniques in Library System

Veepu Uppal
Department of Computer Science and Engineering,
Manav Rachna College of Engineering, Faridabad, India

Gunjan Chindwani
Department of Computer Science and Engineering,
Manav Rachna College of Engineering, Faridabad, India

ABSTRACT
Few years ago, the information flow in library was relatively simple and the application of technology was limited. However, as we progress into a more integrated world where technology has become an integral part of the business processes, the process of transfer of information has become more complicated. Today, one of the biggest challenges that libraries face is the explosive growth of library data and to use this data to improve the quality of managerial decisions. Data mining techniques are analytical tools that can be used to extract meaningful knowledge from large data sets. This paper addresses the applications of data mining in library to extract useful information from the huge data sets and providing analytical tool to view and use this information for decision making processes by taking real life examples.

General Terms
Data mining, Association Rules, Clustering.

Keywords
Library, Classification, Prediction, Outlier analysis, support, confidence.

1. INTRODUCTION
In modern world a huge amount of data is available which can be used effectively to produce vital information. The information achieved can be used in the field of Medical science, Education, Business, Agriculture and so on. As huge amount of data is being collected and stored in the databases, traditional statistical techniques and database management tools are no longer adequate for analyzing this huge amount of data. Data Mining also called data or knowledge discovery has become the area of growing significance because it helps in analyzing data from different perspectives and summarizing it into useful information [1]. How to improve the utilization rate of library resources, how to serve the reader better, and how to play more active roles, all have been becoming the concrete task of library in future. The data mining of the books circulation and user needs in library automation system provided effective support for library management [2]. Many kinds of knowledge can be discovered using data mining on library data. . The discovered knowledge can be used for allocation of books in such a way that circulation of books can be done easily, helps to enhance the interest of readers as well as the loyalty to library. The discovered knowledge can be used to cluster the students or departments according to their behavior and enables the librarian to recognize the departments those he should give special attention. Bibliomining can also be used to predict future user needs. By looking for patterns in high-use items, librarians can better predict the demand for new items in order to determine how many copies of a work to order. The library can observe the frequently borrowed by students. Therefore library should increase number of copies for these books. The discovered knowledge can be used by librarians to look for patterns commonly associated with lost/stolen books and high user fees. Once these patterns have been discovered, appropriate policies can be put in place to reduce inventory losses.

2. DATA MINING DEFINITION AND TECHNIQUES
Simply stated, data mining refers to extracting or "mining" knowledge from large amounts of data. Data mining techniques are used to operate on large volumes of data to discover hidden patterns and relationships helpful in decision making. The steps identified in extracting knowledge from data are:

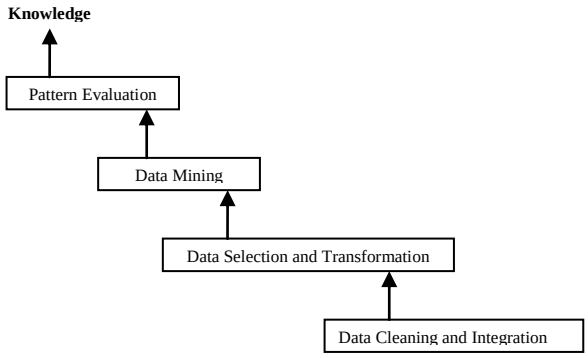

Fig. 1: Steps of extracting knowledge from data

2.1 Association analysis
Association analysis is the discovery of association rules showing attribute-value conditions that occur frequently together in a given set of data. Association analysis is widely used for market basket or transaction data analysis. More formally association rules are of the form $X \Rightarrow Y$, i.e, $A_1 \wedge ---\wedge A_m \Rightarrow B_1 \wedge ---\wedge B_n$

Where A_i and B_j are attribute value pairs.

$i = 1.....m \qquad j = 1.....n$

The association rule is interpreted as database tuples that satisfy the conditions in X are also likely to satisfy the conditions in Y. The rule $X \Rightarrow Y$ holds with support's' if s% of transactions in D contain $X \cup Y$. Rules that have's' greater than a user specified support is said to have minimum support. The rule holds with confidence 'c' if c% of transactions in D that contain X also contain Y. Rules that

have greater 'c' than a user specified confidence is said to have minimum confidence.

2.2 Classification and Prediction

Classification is the processing of finding a set of models which describe and distinguish data classes or concepts, for the purposes of being able to use the model to predict the class of objects whose class label is unknown. The derived model may be represented in various forms, such as classification rules, decision trees, mathematical formulae, or neural networks. Classification can be used for predicting the class label of data objects. However, in many applications, one may like to predict some missing or unavailable data values rather than class labels. This is usually the case when the predicted values are numerical data, and is often specifically referred to as prediction. IF-THEN rules are specified as

IF condition THEN conclusion

Supervised Classification: - The set of possible classes is known in advance.

Unsupervised Classification: - Set of possible classes is not known. After classification we can try to assign a name to that class. Unsupervised classification is called clustering

2.3 Clustering Analysis

Unlike classification and predication, which analyze class-labeled data objects, clustering analyzes data objects without consulting a known class label. In general, the class labels are not present in the training data simply because they are not known to begin with. Clustering can be used to generate such labels. The objects are clustered or grouped based on the principle of maximizing the intraclass similarity and minimizing the interclass similarity i.e. clusters of objects are formed so that objects within a cluster have high similarity in comparison to one another, but are very dissimilar to objects in other clusters. Each cluster that is formed can be viewed as a class of objects, from which rules can be derived. Application of clustering in library can help library to group individual student into classes of similar behavior. Partition the students into clusters, so that students within a cluster are similar to each other while dissimilar to students in other clusters corresponding to issuing the books to students in one cluster.

2.4 Outlier Analysis

A database may contain data objects that do not comply with the general behavior of the data and are called outliers. The analysis of these outliers may help in fraud detection and predicting abnormal values. Causes of outliers are as below:-

- Poor data quality / contamination.
- Low quality measurements, malfunctioning equipment, manual error.
- Correct but exceptional data.

3. POTENTIAL APPLICATIONS

3.1 Predicting the Allocation of Books

It is important for library to allocate the books in such a way that will improve the circulation process and will help the students to search the books. With these strong association rules of the books' classification, the library can have good decision making to achieve optimal allocation of resources and devices. It will have reasonable and better procurement of all types of literature and rational distribution of stock and resources.

Table 1:Books Database

Book ID	Subject
K3	Biography
L1	Chinese literature
Y2	Common foreign languages

The association rule $\mathbf{K3, L1 \Rightarrow Y2}$ **(support: 17.784% and Confidence: 70.777%)** implies that there are more than 70.777% percent readers who borrow books of K3 (biography) and L1 (Chinese literature) at the same time will borrow books of Y2 (common foreign languages), then these books should be arranged on adjacent position while considering these types of books should be convenient for readers to borrow in order to speed up the circulation of books.

The traditional library has been in accordance with the concept of "waiting for readers", and the concept and models will no longer be able to adapt for the service of modern library .After collecting and processing information involving lots of acts of readers in borrowing books, data Mining identified the interests, habits, trends and demand of specific individuals or groups of readers, then inferred the next acts of the corresponding group or individual. So the specific services for the custom content can be supplied for them. This personalized information services not only increase satisfaction of readers and get better use of resources, but also it is conducive to the library' further development compared to the passive service of "waiting for readers". E.g., while readers visiting the new information or bibliographic data which recommended by library in a timely manner, the strong association rules of books will promote related topics information for readers or guide readers to find information through tips in order to give books recommendation and to achieve better personal service. The rule $\mathbf{K3, Y2 \Rightarrow L1}$ **(support: 0.252 %, Confidence: 83.333%)** implies that If library recommended K3 and Y2 when readers borrow I1, the recommendation will enhance the interest of readers as well as the loyalty to library. Although the reader's interest will change along with the development of the times, the library system have the ability to discover the latest needs of readers automatically with the application of data mining technology [2] .

3.2 Protecting Patron Privacy

When an item is returned, many libraries delete all information about that transaction. However, there is valuable decision-making information that is lost. While the operational system is a user-focused data source, the data warehouse is an item-focused data source. Therefore, before deleting the transactional information, a record should be created in the data warehouse that combines information about the item with demographic information about the patron. This will capture the important information about the transaction without identifying the patron involved [4].

Table 2: Original Circulation Record

Book ID	Subject	Patron
QA76.9	Computer Science	392-33
PS159.G8	American Literature	575-49
HF5415.125	Marketing	392-33

Table 3: Original Patron Database

Patron	Name	Class	Dept
373-34	Abhay	Grad	Psych
392-33	Sophie	U.G.	Math
575-49	Kenneth	Faculty	English

Table 4: Data Warehouse Combined Cleaned Circulation Database

Book ID	Subject	Patron Class	Patron Dept
QA76.9	Computer Science	U.G.	Math
PS159.G8	American Literature	Faculty	English
HF5415.125	Marketing	U.G.	Math

3.3 Identifying Departments' Needs

Results of clustering analysis are used to improve service quality and optimize Library Management Models According to the statistical analysis, the departments of Chinese, finance, foreign language and mathematics are the most active departments whose students borrowed much more books than the other departments. But the numbers of students of the above departments are also much more than that of other departments. This can not reflect the objective demand of readers. From the results of cluster analysis based on the average loan rates of departments (book/person/year), we can see that the numbers of some departments is fewer, but the average loan rates of them are relatively high. Then the needs of them should not be ignored, Such as the Fine Arts Department and Music Department whose students are fewer but have the high demand of books, and the depth and breadth of the books purchasing about the art and music should be strengthened. Librarians not only should pay more attention to the specialties and academics research of all subjects, but also should not ignore the needs of some of small departments. Then, the library also can focus on recommending books to the active groups, have interactive communication with readers, and play more active roles to achieve the goal of efficient access to the reader needs and the reasonable books procurement based on the results of clustering [3].

3.4 Patterns Identification for inventory loss

In addition to association rule mining, sequential pattern mining can be applied to extract new useful and interesting patterns from library data in which time parameter is involved. In sequential pattern mining, we take time stamp into account then find the proper rules. By using sequential pattern mining Library can improve its services. The library can observe the frequently borrowed by students. Therefore library should increase number of copies for these books. Sequential pattern mining show student's behavior in borrowing books in the library.

Table 5: Books Circulation Record

Student ID	Book ID	Transaction Date
1	K3	22/11/2010
2	I1	23/11/2010
3	K3	23/11/2010
4	K3	24/11/2010
5	Y2	24/11/2011

By Sequential Pattern Mining patterns commonly associated with lost/stolen books or late returning can be discovered. Once these patterns have been discovered, appropriate policies can be put in place to reduce inventory losses [5].

3.5 Analyses Reader Characteristics by Clustering

The readers of library are clustered in different group according to the grades and major and by the nature of books those they borrow**.**

Table 6:Users Clusters Based on Books Demand

No.	Category	Books	Major
1	Junior	B (Philosophy)	Logistics Management
2	College senior	B (Philosophy)	Marketing
3	Sophomore	Finance	Financial
4	Freshman	I (Literature)& H (Language)	Any

The classification on above data tells us that, the freshman usually borrows category I (Literary) and category H (Language), the distinction of major is not significant. The sophomore, junior and senior are stronger clustered by majors, additionally, junior and senior usually borrow category B (Philosophy) books. It shows that, increasing with level of category, Readers' borrowing trend to more associate with their majors [6].

3.6 Analysis of Book Circulation by Time Series

The data mining on circulation data can be done to do monthly circulation variation tendency analysis. It gives us knowledge about in which month of year more books are circulated. There are two academic semester in a year Feb-May and July-Dec. In the month of Jan there are holidays so the probability of issuing book by students is low. Near the examination in the month of April and May, students issue more books. In the same way June month contains holidays so less books are issued during this period and in the months of November and December more books are issued by students. So, by analyzing the ratio of issued books at regular time series librarian can plan the purchasing, maintenance and enhancements of resources.

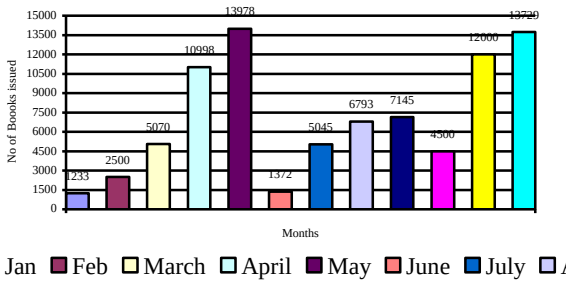

Fig. 2: Books Circulation Analysis by time

3.7 Finding out reader interest by association rules

With the development of information technology, the university library becomes digital library from traditional library gradually. Digitalizing resource and establishing personal recommendation system are main tendency of library advancement. The personal information service is the service that can meet the personal information request of all reader, and can automatically in time offer recommendation information to reader by analyzing user's personality and habit. In order to offer recommendation service, we need get common knowledge such as "various types of readers like what kinds of books". All circulation datum are stored by the way of circulation records in computer, we can find reader's borrow book habits by means of association analysis to circulation records on line, and then establish recommendation model to offer personal information service.

Table 7:Branchwise Book Circulation Record

Branch	Specialty	Book ID
G1	S1	B1,B2,B13
G2	S2	B2,B3
G3	S2	B31,B43
G3	S3	B3,B13,B2

Where specialty number is the intestine code of various specialties in the university, S1 (Undergraduate), S2 (Master Degree Student), S3 (Doctorial Student). (B1, B2, B3, B5, B13, B20, B31, B43) is corresponding to (Electronics, VLSI, Educational technology, Psychology, Computer application, English, Physical training, High mathematics) and Code (G1, G2, G3) is corresponding to (Computer Science, Electronics and communication, Mechanical).

Table 8:Association Rules Generated from Users Interest

No.	Association Rule	Confidence
11	S1,G2 $\Rightarrow$ B1,B2	78%
13	G2,S2 $\Rightarrow$ B2,B3	73%
14	B3 $\Rightarrow$ B13	74%
15	B20 $\Rightarrow$ B43	86%

- Rule11 describes students of computer science and technology specialty in S1 often borrow 'Electronics' and 'VLSI' books in order to meet the need of specialty foundational course study.
- Rule13 describes 73% of students borrowed' VLSI,' Educational technology' books at one time in historical data of Electronics Master Degree Candidates.
- Rule14 represents 74% of students, which borrowed 'Educational technology' books, may borrow 'Computer application' books.
- Rule15 represents 86% of students, which borrowed 'English' books, may borrow 'High mathematics' books, this is mean that the student study 'English' and 'High mathematics' course synchronously [7]

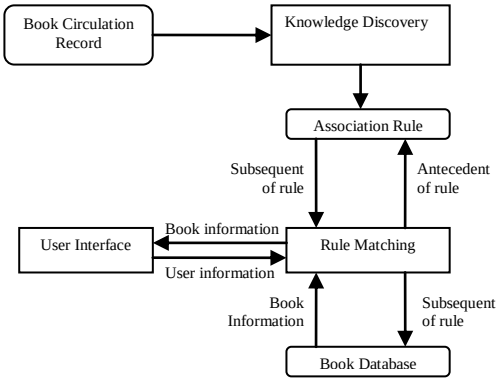

Fig. 3: Model of book recommendation service

4. SUMMARY

This paper discusses the application of association analysis, clustering, and sequential pattern mining on a library dataset to extract frequent book sequences borrowed by students. One of the result shows that readers who borrow books of K3 (biography) and L1 (Chinese literature) at the same time will borrow books of Y2 (common foreign languages). The patterns can be used by library in order to improve its services to students effectively. Number of copies for books occurred in the frequent sequences can be increased to support students in learning related subjects. Library may also give readers recommendations to read other books after they finish reading a certain book. Based on the book occurrences in frequent sequences, layout of books can be arranged such that readers can find easily the books. The model of book recommendation service can offer book information to the readers that may be user-interested in. This model can also be used in other fields, for example, bookstore, information retrieval system, network reference database, etc.

5. REFERENCES

[1] J. Han and M. Kamber, *"Data Mining Concepts and Techniques,"* San Diego, USA: Morgan-Kaufmann, 2006.

[2] JianWei Li and Pinghua Chen *"The Application of Association Rule in Library Syste*m", Knowledge Acquisition and Modeling Workshop, IEEE Conference Publications 2008, pp. 248-251.

[3] JianWei Li and Pinghua Chen *"The Application of Cluster Analysis in Library system"*, Knowledge Acquisition and Modeling Workshop, IEEE Conference Publications 2008, pp. 907 – 910.

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

[5] I. S. Sitanggang et al. *"Sequential Pattern Mining on Library Transaction Data"*, Information Technology (ITSim), IEEE Conference Publications 2010, pp. 1-4.

[6] Ping YU, *"Data Mining in Library Reader Management"* Network Computing and Information Security (NCIS), IEEE Conference Publications 2011, pp. 54 – 57.

[7] Zhen Zhu and Jing-yan Wang *"Book Recommendation Service by Improved Association Rule Mining Algorithm"* Machine Learning and Cybernetics, IEEE Conference Publications 2007, pp. 3864-3869.

[8] Yingying Li and Yimin Zhang "*Application of data mining techniques in sports training*" Biomedical Engineering and Informatics, IEEE Conference Publications 2012, pp. 954 – 958.