Application of Data Mining Techniques for Library Management Information System

By

Bikash Mukhopadhyay

Information Scientist

Central Library

The University of Burdwan

Burdwan , Pin – 713 104

E-mail: buclib@satyam.net.in

Sripati Mukhopadhyay

Professor & Head,

Dept of Computer Science,

The University of Burdwan,

Burdwan – 713 104

E-mail: dgp_uvcompsc@sancharnet.in

ABSTRACT

Data mining is a form of artificial intelligence that uses automated processes to find information. Although its use in libraries is to be explored, data mining has been used successfully for several years in the scientific and business communities for tracking behavior of individuals and groups, processing medical information, and a number of other applications. This system has been designed for librarians to help them manage the library easily. Data Mining involves different kinds of technologies, solutions and techniques and because of its pattern-recognition features; it can be applied to a large database. Data mining software can relate one field to the other field in the database. It can be used to analyze the on-line users' behavior and to predict the future potential users. Here, we have applied data mining to library systems using artificial intelligence technology.

KEY WORDS: Data Mining, Artificial Intelligence, Web Mining, XML, ODBC

0. INTRODUCTION

At the beginning of each adjustment in a library, librarians or library managers need to think about which kind of books are going to purchase or what kind of magazines have to be ordered in the coming renewal period of the year. The purpose of project is to provide statistics and analysis software. It will offer librarians or managers sufficient information to make right decisions. The statistical data or chart produced by the program will easily make librarians sense what changes need to be made. We are going to research the reader’s behavior from these statistical data or charts and then improve or modify our library service. The library professionals do not know which books a patron wants, which is offered for their studying, or which patron prefers what class of book. The library may offer a book that no patron wants to borrow, or the offered books may not have enough readers. Librarians should make a decision so that this kind of book may never be or seldom ordered in the future. On the other hand, librarians have to use a lot of their budget for much used books to provide the reader more services. Librarians will be helped by this program. This kind of feedback can be researched and can present some trends. This will be important information for further improvement of the library. Data Mining in a Library System will be a product after we digitalize the library. Librarians may get a lot of worthwhile information immediately such as the students’ studying trend, the top 10 of circulation in the library, the hottest field of studying by using this system. Not only librarians want this but also patrons need it. This is also a knowledge management strategy system for the current library. Actually, the system will enhance the librarian’s efficiency in controlling library resources. Librarians could imagine the future library from it. The uncertainties of the Internet challenge the modern workplace but also promise unexpected opportunities. Converting uncertainties to opportunities requires human efforts to enhance the capabilities of technology to benefit professional endeavor. In the library profession, practices of unstructured and unrestricted information dissemination and retrieval from the World Wide Web (WWW) not only create an information explosion but make uncertain the future of the profession as well. Facing the need to adapt to the new climate of the information world, academic librarians are charged to extend their services from the reference desk to the virtual environment on the WWW. To meet the challenge, new information technologies that can assist in organizing and retrieving information are constantly being investigated by academic libraries. Web mining is one of these new technologies that deserves special attention because of its capability to discover and analyze useful information from the Web.

1. WHAT ARE DATA MINING AND WEB MINING?

Data mining is one of the hottest topics and buzz word in information technology. It automatically and exhaustively explores very large datasets, consequently uncovering otherwise hidden relationships among data .This technology has been successfully applied in science, health, marketing and finance to aid new discoveries and strengthen markets. In addition, data mining techniques are being applied to discover and organize information from the Web .Data mining itself is in the second generation of artificial intelligence; the core concept of data mining is focused on machine learning .This machine-learning ability allows modification of search criteria automatically before the next execution, once particular patterns or trends have been discovered in the data searched .It is important to understand that data mining is a discovery-oriented data analysis technology and not a single product or a system. It is a highly focused data transformation framework . This transformation process uses a series of analytical techniques, such as clustering, association and classification .These techniques are taken from the field of mathematics, cybernetics and genetics, and can be used independently or cooperatively. The function is to extract high quality information to identify facts and draw conclusions based on relationships or patterns among the data . Most important, data mining can ask a processing engine to "show answers to questions we do not know how to ask" .For example, bank customers' data are kept in different databases, thus, they are isolated from each other. Data mining technology can search all the different databases together, and provide a better customer view so that the bank can concentrate more on potentially good customers .The rationale is that when asking for specific relationships, more important relationships might be missed. Asking to find relationships that we do not know exist will yield more meaningful data or business knowledge .The combination of these two areas, data mining and the WWW, is known as Web mining. When data mining is applied to the Web, it can perform several functions including:

Resource discovery

the discovery of locations of unfamiliar files on the network;

Information extraction

the acquisition of useful information from the WWW;

Generalization

the discovery of information patterns from said resources

There are two primary dimensions of Web mining: Web content mining and Web usage mining.

1.1 Web Content Mining

Web content mining is the "process of information or resource discovery from millions of sources across the World Wide Web " .There are two approaches in Web content mining: the agent-based and database approaches. The agent-based approach involves artificial intelligence systems that can "act autonomously or semi-autonomously on behalf of a particular user, to discover and organize Web-based information " .Some intelligent Web agents can use a user profile to search for relevant information, then organize and interpret the discovered information. Some use various information retrieval techniques and the characteristics of open hypertext documents to organize and filter retrieved information. Another kind of agent is programmed to learn user preferences and use those preferences to discover information sources for those particular users. The database approach focuses on "integrating and organizing the heterogeneous and semi-structured data on the Web into more structured and high-level collections of resources." These organized resources can then be accessed and analyzed . These "metadata, or generalizations, are then organized into structured collections (e.g., relational or object-oriented databases) and can be analyzed".

1.2 Web Usage Mining

The other dimension of data mining is Web usage mining. This is the process of discovering user access patterns (or user habits), as data are automatically collected in daily access logs. Recently, referrer logs, which collect information about referring pages for each reference and user registration, also have been included. Web usage mining is crucial in establishing user profiles for a better structured Web site. "As the manner in which the Web is used continues to expand, there is a continual need to figure out new kinds of knowledge about user behavior that needs to be mined for" .

2. TEN STEPS OF DATA MINING

The following are ten steps of Data Mining for extracting hidden knowledge from data:

2.1 Identify the Objective

The purpose of Step 1 is creating a list of analyses to perform, questions to investigate, and hypotheses to test. The first of thing is to be clear on what you hope to achieve with your analysis. It includes some advanced planning about what type and level of information you intend to capture from the database. Make sure whether or not the objective is measurable. This advance planning can

save you time and money in your data mining efforts.

2.2 Select the Data

The next step is to select the data to meet this objective. In this step, the data is organized for analysis and to produce an effective representation. The following are some examples for a quick checklist:

Is the data adequate to describe the phenomena, the data mining analysis is attempting to model?

Is there a common field in your data that can be used for linking it to other database?

Are internal and external information available for the analysis?

Is the data stable—will the data being mined be the same and available after the analysis?

Is there redundancy in the data sets when they are merged?

How current and relevant are the data to the business objective?

Is there lifestyle or demographic data available?

2.3 Prepare the Data

Once you've organized the data, you must decide which features to convert into usable formats. The purpose of this step is to produce a “data mining-ready” data set.

Establish strategies for handling missing data, extraneous noise, and outliers

Identify redundant variables in the dataset and decide which fields to exclude

Determine the distribution frequencies of the data

For example, it may be efficient to have an account establish date in the format of YY-MM-DD, but it may be necessary to transform this field to one total account number days like NNNN. You can postpone some of these decisions until you select a data mining tool. For example , if you need a neural network or polynomial network you may have to transform some of your fields.

2.4 Evaluate the Data

“Evaluate the structure of your data in order to choose the appropriate tools.

What is the ratio of categorical/binary attributes in the database?

What is the nature and structure of the database?

What is the overall condition of the dataset?

What is the distribution of the data set?

Balance the objective evaluation of the structure of your data against your users' need to understand the findings.” For example, neural networks generally work best on data sets with a vast number of numeric attributes.

2.5 Format the Solution

In conjunction with the evaluation of data and business objective determine the format of the solution. You may think about

What is the optimum format of the solution—decision tree, rules, C code, graph, or SQL syntax?

What are the available format options from the data mining process?

What is the goal of the solution, classification, and segmentation?

What does the end-client need—graphs, reports, code?

2.6 Select the Tools

This step is to select an appropriate data mining tool for the business objectives and data structure. While selecting a data mining tool, we may consider

Ø Is the data set heavily categorical?

Ø What platforms do your tools support?

Ø Are the tools ODBC-compliant and incorporated with an end-user interface?

Ø What data formats can the tools import/export?

A data mining tool should help you understand the results of its analysis and the kind of solutions the tool generates. No single tool is able to provide the answer to your data mining project. Some tools integrate several technologies into a suite of statistical analysis programs, a neural network, and a symbolic classifier.

2.7 Construct the Model

At this stage the data mining process begins. “This is the process of searching for patterns in a data set and the generation of classification rules, decision trees, clustering, scores, weight, and evaluation and comparison of error rates. Usually the first step is to use a random number seed to split the data into a training set and a test set and construct and evaluate a model.” Resolve these issues:

Ø What are the model error rates? Are they at acceptable levels? Can they be improved?

Ø What extraneous attributes did you find? Can those noisy and redundant be purged?

Ø Is additional data or a different methodology necessary to improve model performance?

Ø Will you have to train and test a new data set?

2.8 Validate the Findings

This step involves testing a model. It is important after achieving your data mining analysis that you share and discuss the results of the analysis with the administrator, designer, analysts, managers, and engineers. Ensure that the findings are correct and appropriate to the objectives.

Are the findings available and do they make sense?

Does the process have to return to any previous steps to improve results?

Can other data mining tools be used to produce the same findings?

2.9 Deliver the Findings

This step provides a final report to the business unit or client. The report should document the entire data mining process such as data preparation, tools used, test results, source code, and rules. Thus, you will know if the findings meet the objective, if additional data can improve the analysis, what strategic insight you dig out and how it is applicable, and what proposals can result from the data mining analysis.

2.10 Integrate the Solutions

Finally, you share the findings with all interested end-users in the appropriate business units. This process involves incorporating the results of the analysis into the company's business procedures. Although data mining tools automatically invoke database analysis, they can make some mistaken findings and erroneous conclusions if you're not careful.

3. ADVANTAGES OF DATA MINING

Data mining uses advanced technology for gleaning valuable insights from database that enable the business user to make the right business decisions. It turns up data that gives business companies just a little edge, but the payoff for this small edge is large. One obtains the competitive advantages required to grow in today's competitive environment. Data mining is simple in theory, but it can get quite involved in practice. That means that users will probably need some help getting set up. Many data mining software solutions are available for small and large businesses in most industries and these tools are constantly improving. The accuracy of mining will make a difference in the bottom line of business. "A recent META Group survey revealed that Fortune 500 firms were using data mining for three general purposes: 64 percent for strategic planning, 49 percent for competitive intelligence, and 46 percent to increase their market share." So far, the following have been the traditional applications of data mining technology by industry sectors:

Retailers Database marketing, advertising effectiveness, inventory and category management.

Banking and Insurance Financial modeling, fraud detection, market and industry profiling, database marketing, customer segmentation.

Brokerage/Stock Exchange

Telecommunication call detail record analysis, optimal use of capital equipment, targeted marketing.

Government Collections, workload selection, logistics, and intelligence gathering.

Manufacturing Production quality control, supply-chain and inventory control

Data mining provides companies extremely valuable insight that has, to date, been closely guarded as corporate stealth technology to protect competitive advantage. However, there is sometimes a breakthrough when trying something new. Some mistakes will be made. Mistakes can be remedied with statistical expertise, but the technology is becoming more useful for more users. Data mining is also not designed to fix all business problems or tell you what the real problem is, but it gives a company a sense which can be used to help make intelligent business decisions.

"Most of all, data mining is an ongoing process that involves a lot of analysis and refining along the way, so think of it as a worthy investment. And like any investment, even if your data-mining portfolio only uncovers small golden nuggets at first, those nuggets properly managed yield a lot of value."

4. DATA MINING TECHNIQUES ENHANCE LIBRARY SYSTEM:

Suppose a library system has the following goals:

1. Increase books borrowing rate.
2. Attract more users to borrow books.
3. Assist library professionals in making policy on the acquisition of duplicate copies and new publications.

The relationship between Data Mining and Library Management can be shown as in figure-1

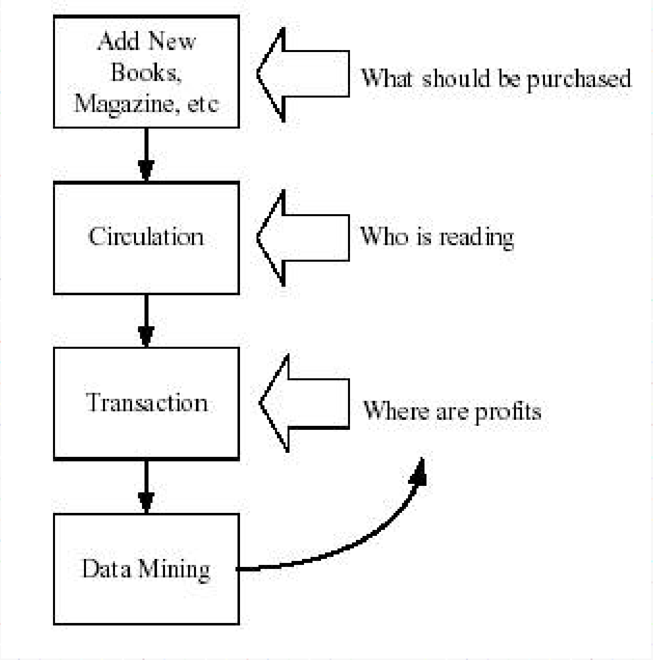

Figure – 1, The cyclical relationship between data mining and library management

Example:

Consider a library whose books database includes the data of 2000 volumes of books , and the borrowing history of 5000 users who have borrowed books from library. From this database, we have developed a report system to represent the relationship between books and students. And then we use the data mining association rule to analyze the books of the same cluster borrowed by the students. A report chart of books will be produced, so that the library professionals will find references to the books for making a decision. Additionally, the library collection may be used more effectively.

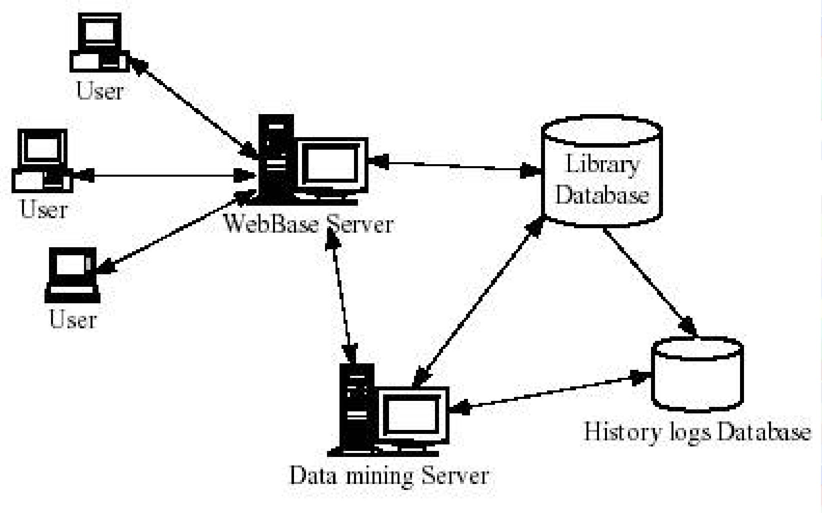

Figure – 2 ,   The prototype of a Library architecture

The library database includes seven tables. The first of them is *BOOKS* which contains each of the books' basic information. The table *BOOKS* was improved from *MARCBOOKS* which comes from the previous library system. It kept most of the columns from MARCBOOKS. Machine-readable Cataloging (MARC) is a kind of metadata to describe a book. *USERS* has part of a student's basic information from the registration system and coordinates with the library system. *COPY* describes the actual volumes in the library. *LOCATION* reports the real location of each book. *CIRCUL* records each circulation of a book. *RESERVATION* has the information that is stored when a user holds a book. Figure - 3 shows the schema of these tables. Between each table, there is a key column that sometimes is a primary key in one of them. For instance, the relation key between *BOOKS* and *COPY* is ISBN.

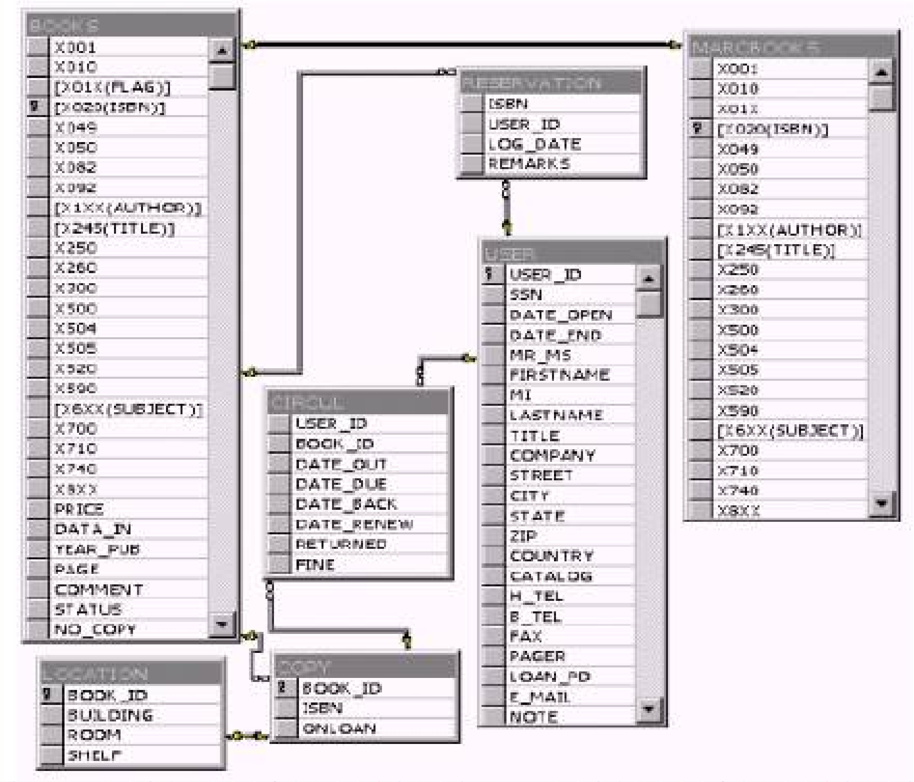

Figure – 3 , The schema of the library database

The data mining system is under statistics in a library system. It is an advantageous part of the library system. The process between the main page of data mining and its four functions is shown in figure- 4. On the main page of data mining, the administrator has four functions to execute. The first function is backup the *CIRCUL*. The second one is the list of top ten books. The third one is to execute rule. The final one is rule manager. The following paragraphs are describing them.

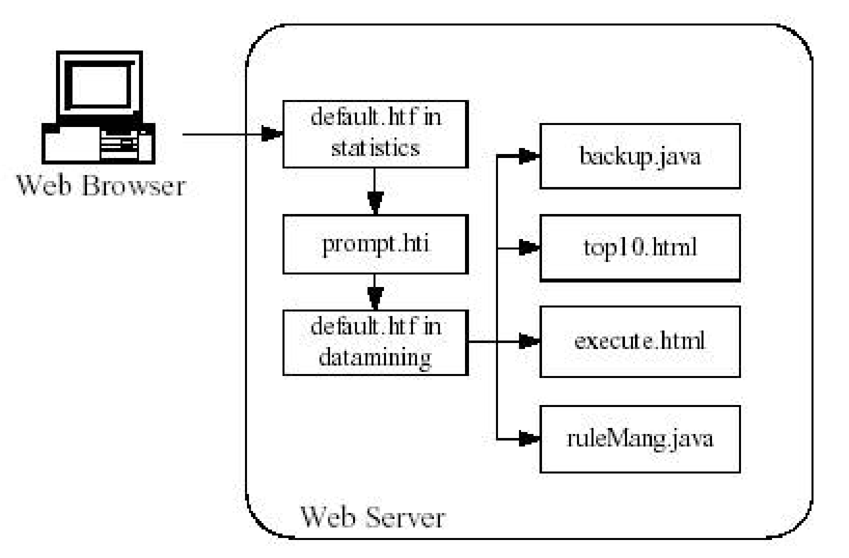

Figure – 4 ,   The main process of the data mining system.

5.   CONCLUSION

It is unlikely that libraries will continue to satisfy users' information needs using the traditional catalogs as the primary access mechanisms. To be certain, online catalogs provide good access to books, films, microfiche, audio tapes, and other materials traditionally kept in libraries. However, few, if any, libraries have succeeded in using their online catalogs to provide adequate access to a significant number of digital materials. In an era where information costs rapidly increase while budgets remain flat, libraries must find alternatives to slow, awkward, and expensive manual cataloging. The Artificial Intelligence methods such as genetic algorithms and machine learning, or statistics method support data mining tools to accomplish tasks. These methods are fast and thorough and exploit the tremendous power of computers to find nuggets of useful information embedded in mountains of data. In the future, advanced data mining technology will depend on computer architecture and database technology.

REFERENCES

[1] Pujari, Arun, K. "Data Mining Techniques" Hyderabad: Universities Press

[2] Richardson, John V. "Knowledge –Based System for General Reference Work " Academic Press, P 289-303.

[3] Mena, Jesus. "Data Mining your Website " MA: Digital Press ,1999.

[4] Krimberly, Kowal "The Library of Congress Classification System (LCC)" <http://library.tulane.edu/lc.htm (Sept,1999)

[5] Gilman, Michael."Data Mining Overview" < http://www.data-mine.com (May, 2000)

{6] WebBase, Inc. < http://www.webbase.com

BRIEF BIOGRAPHY OF AUTHORS

Bikash Mukhopadhyay holds M C A from North Bengal University and presently working as Information Scientist at Burdwan University. His area of interests are Content Management, Data Mining , Knowledge Based Computing Systems and currently doing Ph.D. He has publications in national journals conferences. He is a life member of Bengal Library Association and IASLIC

Prof. Sripati Mukhopadhyay is the Professor and Head, Department of Computer Science, Burdwan University . He received M.Tech in Computer Science and Ph.D from IIT, Kharagpur. His research areas are in the field of Artificial Intelligence, Discrete Maths & Cryptography, Knowledge Base Computing System and Data Mining. He has many publications in international and national journals and proceedings of national and international conferences. He is a life member of Computer Society of India