Application of Data Mining in Library and Information Services

K Prakash Prem Chand Umesh Gohel

Abstract

Knowledge Discovery or Data Mining is the partially automated process of extracting patterns, usually from large data sets. Library and information services in schools, colleges, universities, corporations and communities obtain information about their users, circulation history, resources in the collection, and search patterns. Now a days many libraries have taken advantage of these data as a way to improve customer service, manage acquisition budgets, or influence strategic decision-making about uses of information in their organizations. The paper tries give an overview on data sources and possible applications of data mining techniques in the library.

Keywords : Bibliomining, Data Mining, Knowledge Management.

0. Introduction

We live in the Age of Information. The importance of collecting data that reflect in business or scientific activities to achieve competitive advantage is widely recognized now. Powerful systems for collecting data and managing it in large databases are in place in all large and mid-range institutions and organizations. However, the bottleneck of turning this data into success is the difficulty of extracting knowledge about the system we study from the collected data to mention few here;

- How large are the peak loads of a telephone or energy network going to be ?
- What is the peak time for charging/discharging of books in circulation counter ?
- What goods should be promoted to this customer ?
- Will this customer default on a loan or pay back on schedule ?
- What medical diagnosis should be assigned to this patient ?

These are all the questions that can probably be answered if information hidden among megabytes of data in database can be found explicitly and utilized. Modeling the investigated system, discovering relations that connect variables in a database are the subjects of data mining. Modern computer data mining systems self learn from the previous history of the investigated system, formulating and testing hypotheses about the rules which this system obeys. When concise and valuable knowledge about the system of interest has been discovered, it can and should be incorporated into some decision support system which helps the manager to make wise and informed business decisions.

1. Data, Information, Knowledge and Data Mining

1.1 Data

Data are any facts, numbers, or text that can be processed by a computer. Today, organizations are accumulating vast and growing amounts of data in different formats and different databases.

This includes:

- operational or transactional data such as, sales, cost, inventory, payroll, and accounting
- nonoperational data, such as industry sales, forecast data, and macro economic data
- meta data - data about the data itself, such as logical database design or data dictionary definitions.

1.2 Data Format

Data items can exist in many formats such as text, integer and floating-point decimal. Data format refers to the form of the data in the database.

1.3 Information

The patterns, associations, or relationships among all this *data* can provide *information*. For example, analysis of retail point of sale transaction data can yield information on which products are selling and when.

1.4 Knowledge

Information can be converted into *knowledge* about historical patterns and future trends. For example, summary information on retail supermarket sales can be analyzed in light of promotional efforts to provide knowledge of consumer buying behavior. Thus, a manufacturer or retailer could determine which items are most susceptible to promotional efforts.

1.5 Binning

A data preparation activity that converts continuous data to discrete data by replacing a value from a continuous range with a bin identifier, where each bin represents a range of values. For example, age could be converted to bins such as 20 or under, 21-40, 41-65 and over 65.

1.6 Data Mining

Data mining can be defines as “An information extraction activity whose goal is to discover hidden facts contained in databases. Using a combination of machine learning, statistical analysis, modeling techniques and database technology, data mining finds patterns and subtle relationships in data and infers rules that allow the prediction of future results. Typical applications include market segmentation, customer profiling, fraud detection, evaluation of retail promotions, and credit risk analysis.”

1.7 Bibliomining

Use of data mining to examine library data records might be aptly termed bibliomining. With widespread adoption of computerized catalogs and search facilities over the past quarter century, library and information scientists have often used bibliometric methods (e.g. the discovery of patterns in authorship and citation within a field) to explore patterns in bibliographic information. During the same period, various researchers have developed and tested data mining techniques — advanced statistical and visualization methods to locate non-trivial patterns in large data sets. Bibliomining refers to the use of these techniques to plumb the enormous quantities of data generated by the typical automated library.

2. Reasons for the growing popularity of Data Mining

2.1 Growing Data Volume

The main reason for necessity of automated computer systems for intelligent data analysis is the enormous volume of existing and newly appearing data that require processing. The amount of data accumulated each day by various businesses, scientific, and governmental organizations around the world is daunting. According to information from GTE research center, only scientific organizations store each day about 1 TB (terabyte!) of new information. And it is well known that academic world is by far not the leading supplier of new data. It becomes impossible for human analysts to cope with such overwhelming amounts of data.

2.2 Limitations of Human Analysis

Two other problems that surface when human analysts process data are the inadequacy of the human brain when searching for complex multifactor dependencies in data, and the lack of objectiveness in such an analysis. A human expert is always a hostage of the previous experience of investigating other systems. Sometimes this helps, sometimes this hurts, but it is almost impossible to get rid of this fact.

2.3 Low Cost of Machine Learning

One additional benefit of using automated data mining systems is that this process has a much lower cost than hiring an army of highly trained (and paid) professional statisticians. While data mining does not eliminate human participation in solving the task completely, it significantly simplifies the job and allows an analyst who is not a professional in statistics and programming to manage the process of extracting knowledge from data.

3. Tasks Solved by Data Mining

1. **Predicting** - The task of learning a pattern from examples and using the developed model to predict future values of the target variable.
2. **Classification** - The task of finding a function that maps records into one of several discrete classes.
3. **Detection of relations** - The task of searching for the most influential independent variables for a selected target variable.
4. **Explicit modeling** - The task of finding explicit formulae describing dependencies between various variables.
5. **Clustering** - The task of identifying groups of records that are similar between themselves but different from the rest of the data. Often, the variables providing the best clustering should be identified as well.
6. **Market Basket Analysis** - Processing transactional data in order to find those groups of products that are sold together well. One also searches for directed association rules identifying the best product to be offered with a current selection of purchased products.
7. **Deviation Detection -** The task of determining the most significant changes in some key measures of data from previous or expected values.

4. Extraction Methods

When the data is large and the computations are complex, data mining can be thought of as algorithms for executing very complex queries on non-main-memory data.

4.1 Stages of the Data-Mining Process

1. Data gathering, e.g., data warehousing, Web crawling.
2. Data cleansing: eliminate errors and/or bogus data, e.g., patient fever = 125.
3. Feature extraction: obtaining only the interesting attributes of the data, e.g., \date acquired" is probably not useful for clustering celestial objects, as in Skycat.
4. Pattern extraction and discovery. This is the stage that is often thought of as \data mining," and is where we shall concentrate our e_ort.
5. Visualization of the data.

6. Evaluation of results; not every discovered fact is useful, or even true! Judgement is necessary before following your software's conclusions.

5. How does data mining work ?

While large-scale information technology has been evolving separate transaction and analytical systems, data mining provides the link between the two. Data mining software analyzes relationships and patterns in stored transaction data based on open-ended user queries. Several types of analytical software are available: statistical, machine learning, and neural networks. Generally, any of four types of relationships are sought:

Classes :
Stored data is used to locate data in predetermined groups. For example, a restaurant chain could mine customer purchase data to determine when customers visit and what they typically order. This information could be used to increase traffic by having daily specials.

Clusters :
Data items are grouped according to logical relationships or consumer preferences. For example, data can be mined to identify market segments or consumer affinities.

Associations :
Data can be mined to identify associations. The beer-diaper example is an example of associative mining.

Sequential patterns :
Data is mined to anticipate behavior patterns and trends. For example, an outdoor equipment retailer could predict the likelihood of a backpack being purchased based on a consumer's purchase of sleeping bags and hiking shoes.

5.1 Data mining consists of five major elements

1. Extract, transform, and load transaction data onto the data warehouse system.
2. Store and manage the data in a multidimensional database system.
3. Provide data access to business analysts and information technology professionals.
4. Analyze the data by application software.
5. Present the data in a useful format, such as a graph or table.

5.2 Different levels of analysis are available

Artificial neural networks :
Non-linear predictive models that learn through training and resemble biological neural networks in structure.

Genetic algorithms :
Optimization techniques that use processes such as genetic combination, mutation, and natural selection in a design based on the concepts of natural evolution.

Decision trees :
Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees

(CART) and Chi Square Automatic Interaction Detection (CHAID). CART and CHAID are decision tree techniques used for classification of a dataset. They provide a set of rules that you can apply to a new (unclassified) dataset to predict which records will have a given outcome. CART segments a dataset by creating 2-way splits while CHAID segments using chi square tests to create multi-way splits. CART typically requires less data preparation than CHAID.

Nearest neighbor method :
A technique that classifies each record in a dataset based on a combination of the classes of the *k* record(s) most similar to it in a historical dataset (where *k* 1). Sometimes called the *k*-nearest neighbor technique.

Rule induction :
The extraction of useful if-then rules from data based on statistical significance.

Data visualization :
The visual interpretation of complex relationships in multidimensional data. Graphics tools are used to illustrate data relationships.

6. What technological infrastructure is required ?

Today, data mining applications are available on all size systems for mainframe, client/server, and PC platforms. Enterprise-wide applications generally range in size from 10 gigabytes to over 11 terabytes. There are two critical technological drivers:

Size of the database :
The more data being processed and maintained, the more powerful the system required.

Query complexity :
Tthe more complex the queries and the greater the number of queries being processed, the more powerful the system required.

Relational database storage and management technology is adequate for many data mining applications less than 50 gigabytes. However, this infrastructure needs to be significantly enhanced to support larger applications. Some vendors have added extensive indexing capabilities to improve query performance. Others use new hardware architectures such as Massively Parallel Processors (MPP) to achieve order-of-magnitude improvements in query time. For example, MPP systems from NCR link hundreds of high-speed Pentium processors to achieve performance levels exceeding those of the largest supercomputers.

7. Applications

A data mining application is an implementation of data mining technology that solves a specific business or research problem. Example application areas include:

1. Decision trees constructed from bank-loan histories to produce algorithms to decide whether to grant a loan.
2. Patterns of traveler behavior mined to manage the sale of discounted seats on planes, rooms in hotels, etc.
3. Skycat and Sloan Sky Survey: clustering sky objects by their radiation levels in different bands allowed astromomers to distinguish between galaxies, nearby stars, and many other kinds of celestial objects.

4. Comparison of the genotype of people with/without a condition allowed the discovery of a set of genes that together account for many cases of diabetes. This sort of mining will become much more important as the human genome is constructed.

5. A pharmaceutical company can analyze its recent sales force activity and their results to improve targeting of high-value physicians and determine which marketing activities will have the greatest impact in the next few months. The data needs to include competitor market activity as well as information about the local health care systems. The results can be distributed to the sales force via a wide-area network that enables the representatives to review the recommendations from the perspective of the key attributes in the decision process. The ongoing, dynamic analysis of the data warehouse allows best practices from throughout the organization to be applied in specific sales situations.

6. A credit card company can leverage its vast warehouse of customer transaction data to identify customers most likely to be interested in a new credit product. Using a small test mailing, the attributes of customers with an affinity for the product can be identified. Recent projects have indicated more than a 20-fold decrease in costs for targeted mailing campaigns over conventional approaches.

7. A diversified transportation company with a large direct sales force can apply data mining to identify the best prospects for its services. Using data mining to analyze its own customer experience, this company can build a unique segmentation identifying the attributes of high-value prospects. Applying this segmentation to a general business database such as those provided by Dun & Bradstreet can yield a prioritized list of prospects by region.

8. A large consumer package goods company can apply data mining to improve its sales process to retailers. Data from consumer panels, shipments, and competitor activity can be applied to understand the reasons for brand and store switching. Through this analysis, the manufacturer can select promotional strategies that best reach their target customer segments.

8. Applications in Library and Information Service

Most people think of libraries as the little brick building in the heart of their community or the big brick building in the center of a campus. These notions greatly oversimplify the world of libraries, however. Most large commercial organizations have dedicated in-house library operations, as do schools, non-governmental organizations, as well as local, state, and central governments. With the increasing use of the Internet and the World Wide Web, digital libraries have proliferate, and these serve a huge variety of different user audiences, e.g., people interested in health and medicine, science and technology, industry and world news, law, and business. With this expanded view of libraries, two key insights arise. First, libraries are nearly always embedded within larger institutions. Corporate libraries serve their corporations, academic libraries serve their universities, and public libraries serve the general public. Second, libraries play a important role within their institutions as repositories and providers of information resources. In the provider role, libraries represent in microcosm the intellectual, learning, and knowledge management activities of the people who comprise the institution.

This fact provides the basis for the strategic importance of library data mining: By ascertaining what users need to know and how well those needs are served, bibliomining can reveal insights that have meaning in the context of the library's host institution. Using bibliomining, libraries can ascertain what their constituencies want to learn, whether they find the information they seek, and whether that information satisfies their learning and knowledge needs. In corporate libraries, which serve the knowledge needs of commercial organizations, such insights can help to develop and maintain a competitive, cutting edge

workforce. In special libraries, which support the research needs of government and non-governmental organizations, these insights can influence the success of policies and programs that are informed by research. In academic organizations, accurate insights into faculty and student knowledge needs can enhance the viability of the whole institution.

To understand libraries can help to achieve these insights, and thus help to enhance the effectiveness of their host organizations or communities, it is important to understand the workflow and associated dataflows that occur within a prototypical library.

8.1 Overview of Library Workflow

Workflow in a traditional “bricks and mortar” library creates a number of data sources appropriate for bibliomining. Before a library obtains new information resources (e.g., books, databases, reference tools, electronic access, etc.), a librarian assesses the needs of the existing collection in light of available and upcoming publications. Next, acquisitions personnel obtain the information resources specified from this needs assessment. Once the library obtains requested new resources, cataloging personnel either create or purchase a catalog record for the new resource. The circulation department then makes the resource available to end-users. Depending upon the size of the library and the scope of its operations, these activities fall within the purview of one, a dozen, or possibly hundreds of different employees organized into specialized departments.

After an information resource appears in the library’s collection, users locate it using catalog search systems and bibliographic databases. Although little uniformity exists with regard to the specifics of the user interfaces for these systems, most catalogs and bibliographic databases support a standard Web browser client as the user interface. Increasingly, catalogs and databases are cross-linked, and each user’s search record and traversal of links appears in log files. When users find resources that they wish to borrow, the circulation department records their selection in a database that tracks the location of each resource owned by the library. As this overview suggests, all functional processes of the library – collection assessment, acquisition, cataloging, end user searching, and circulation – generate large reserves of available data that document information resource acquisition and use. Library information systems frequently use large relational databases to store user information, resource information, circulation information, and possibly bibliographic search logs.

The vast data stored in the databases of traditional and digital libraries represent the behavioral patterns of two important constituencies: library staff and library users. In the case of library staff, mining available acquisitions and bibliographic data could provide important clues to understanding and enhancing the effectiveness of the library’s own internal functions. Mining user data for knowledge about what information library users are seeking, whether they find what they need, and whether their questions are answered, could provide critical insights useful in customer relations and knowledge management. These kinds of information can have strategic utility within the larger organization in which the library is situated.

Integrated Library Systems and Data Warehouses. Although the system used in most parts of the library is commonly known as an Integrated Library System (ILS), very few ILS vendors facilitate access to the data generated by the system in an integrated fashion. Instead, most librarians conceptualize their system as a set of separate data sources. While a relational database stands at the heart of most ILS systems, few system vendors provide sophisticated analytical tools that would promote useful access to this raw data. Instead, vendors encourage library staff to use pre-built front ends to access their ILS databases; these front ends typically have no features that allow exploration of patterns or findings across multiple data sets. As a first step, most managers who wish to explore bibliomining will need to work with the technical staff of their ILS vendors to gain access to the databases that underlie their system.

Once the vendor has revealed the location and format of key databases, the next step in bibliomining is the creation of a data warehouse. As with most data mining tasks, the cleaning and pre-processing of the data can absorb a significant amount of time and effort. A truly useful data warehouse requires integration methods to permit queries and joins across multiple heterogeneous data sources. Only by combining and linking different data sources can managers uncover the hidden patterns that can help understand library operations and users. After the data warehouse is set up, it can be used for not only traditional SQL-based question-answering, but also online analytical processing (OLAP) and data mining. Multidimensional analysis tools for OLAP (e.g., Cognos) would allow library managers to explore their traditional frequency-based data in new ways by looking at statistics along easily changeable dimensions. The same data warehouse that supports OLAP also sets the stage for data mining. This data warehouse will lower the cost of each bibliomining project, which will improve the cost/benefit ratio for these projects. The remainder of this chapter builds on the assumption that this data warehouse is available.

8.2 Bibliomining to Improve Library Services

The users of library services are one of the most important constituencies in most library organizations. Most libraries exist to serve the information needs of users, and therefore, understanding those needs is crucial to a library's success. Examining individual users' behaviors may aid in understanding that individual, but it tells librarians very little about the larger audience of users. Examining the behaviors of a large group of users for regular patterns can allow the library to have a better idea of the information needs of their user base, and therefore better customize the library services to meet those needs.

For many decades libraries have provided readers' advisory services with the help of librarians who know the collection well enough to help a user choose a work similar to other works. Market analysis can provide the same function by examining circulation histories to locate related works. In addition, this information could be provided to the OPAC to allow users to see similar works to one they have selected based upon circulation histories. While it is technically possible to build a profile for users based upon their own circulation history (Amazon.com for example), it may be legally and ethically questionable to do this without a user's permission. Nonetheless, by obtaining and using anonymous data from a large number of users, one can obtain similar results.

In order to locate works in the library, users rely on the OPAC. Librarians often examine user comments and surveys to assess user satisfaction with these tools. Therefore, librarians may wish to examine the artifacts of those searches for problem areas instead of relying on user comments and surveys in order to improve the user experience. When upgrading or changing library system interfaces, librarians can explore these patterns of common mistakes in order to make informed decisions about system improvements.

8.3 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. To prevent inventory loss, predictive modeling can be used 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. In addition, fraud models can be used to determine the appropriate course of action for users who are chronically late in returning materials. The library can also better serve their user audience by determining areas of deficiency in the collection. The reference desk and the OPAC are two sources of data that can aid in solving problems with the collection. If the topics of questions asked at the reference desk are recorded along with the perceived outcome of the interaction, then patterns can be discovered to guide librarians to areas that need attention in the collection.

9. Conclusion

Owing to automation of libraries, they have gathered data about their collections and users for years, but have rarely used those data for better decision-making. By taking a more active approach based on applications of data mining, data visualization, and statistics, information organizations can get a clearer picture of their information delivery and management needs. At the same time, libraries must continue to protect their users and employees from misuse of personally identifiable data records. Libraries must compete against online booksellers, downloadable audio books, and the vast supply of "free" information of varying quality from the Internet, librarians must begin to take the initiative in using their systems and data for competitive advantage and to justify continued support and funding of libraries. The process of using library data more effectively begins by discovering ways to connect the disparate sources of data most libraries create. Connecting these disparate sources in data warehouses can facilitate systematic exploration with different tools to discover behavioral patterns of the libraries primary constituencies. These patterns can help enhance the library experience for the user, can assist library management in making decisions and setting policies, and can assist the parent organizations or communities to understand information needs of their members.

10. References

1. Banerjee, K. (1998). Is data mining right for your library ? Computers in Libraries, vol. 18(10), 28-31.
2. Chaudhry, A. S. (1993). Automation systems as tools of use studies and management information. IFLA Journal, vol. 19(4), 397-409.
3. Data Mining: What is Data Mining?http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm. Visited on September 2004
4. Data Mining. http://www.ibm.com/sfasp/locations/milan/index.html. Visited on September 2004.
5. Guenther, K. (2000). Applying data mining principles to library data collection. Computers in Libraries, 20(4), 60-63.
6. Nicholson, S and Stanton, J. Gaining Strategic Advantage through Bibliomining: Data Mining for Management Decisions in Corporate, Special, Digital, and Traditional Libraries. http://www.scottnicholson.com. Visited on September 2004.
7. Zhang C.; Wang P.; Zhao Y.; Lai Q.; Kong L.(2003) Network information resources management system based on knowledge mining. Online Information Review, vol. 27, no. 2, pp. 129-135(7)

About Authors

Mr. K Prakash is working as Scientific/Technical Officer-I with INFLIBNET Centre since 1995. He has his basic degree in Science and Masters Degree in Library and Information Science from Karnatak University, Dharwad. He has qualified SLET. Pursuing research in Library Automation. He has done specialization course in "Information Technology Applications to Library and Information Services" from NCSI, IISc Bangalore. Before joining to INFLIBNET, he has worked in academic and industrial libraries. Presently he is working in Serials Union Database Development & Managing, and in addition to this he is involved in Training and other Activities of the centre. He has contributed several papers in Seminars and conferences. He is a life member of several professional bodies and he is managing digilib_India forum also. His areas of interests are Library Automation, Database Management, Information Retrieval, Organisation of e-resources, Digital Libraries and Training etc.

Mr. Prem Chand started his career in the year 1990 from Lal Bahadur Shastri National Academy of Administration, Mussoorie. Joined INFLIBNET in 1996 and has around 13 years of experience in Library Automation and Networking. He has published many papers in various journals and conferences. He has completed number of projects. Presently engaged in Development and Maintenance of online union catalogue of different types of document of university libraries in India. His areas of interest are Bibliographic Standards, ILL, Library Consortia , Library Automation and Networking.

Mr. Umesh Gohel has more than nine years of experience in system analysis, design, development, testing and implementation of softwares in various hardware and software platforms as Member of Database R & D Group, INFLIBNET. He is having expertise in Web-Development environment using J2EE, WebLogic Server, PERL and Sybase Adaptive Server 11.5 and Ms-SQL Server 6.5/7.0/2000 as backend Database Server. He had efficiently led the project of Search Engine development for Union Catalogue, which is online at INFLIBNET Servers. He is responsible for administration and maintenance of Sybase Adaptive Server(s) of Union Catalogue Databases. He has also contributed substantially in the development of SOUL (Software for University Libraries), specifically in ISO-2709 to SOUL and vise versa, Circulation and OPAC Module. He has also developed many interface software based CDS/Pascal for Database Authentication.