Data Mining and its use in Libraries

By

Anil Kumar Dhiman

Central Library

Gurukul Kangri University

Haridwar-249 404 (U.A.)

ABSTRACT

Data mining is defined as the automatic extraction of information from data warehouses, enabling an organization to improve its performance by getting new opportunities. It is relatively new term in the world of library science though it is being used in business organization since a long time. This paper gives an overview of data mining and its use in the field of library science.

KEY WORDS: Data Warehouses, Data Mining, Kdd And Libraries.

0. INTRODUCTION

Data Mining is currently regarded as the key element of a much more elaborated process called Knowledge Discovery in Databases (KDD). The knowledge is a collection of interesting and useful patterns in a database, whereas, KDD in a database is the non-trivial extraction of implicit, previously unknown and potentially useful information from data. The knowledge is stored in data warehouse, which is the central storehouse of data that has been extracted from operational data over a time in a separate database. The information in a data warehouse is subject oriented, non-volatile and of an historic nature, so they contain extremely large datasets. Data mining is the automatic extraction of patterns of information from these historical data or so called data warehouses, enabling organizations to focus on the next important aspects of their business -telling them what they did know and had not even thought of asking. It is related to the sub-areas of statistics called exploratory data analysis, which has similar goals and relies on statistical measures; also closely related to the sub areas of artificial intelligence called knowledge discovery and machine learning. The important distinguishing characteristic of data mining is that the volume of data is very large, although ideas from these related areas of study are applicable to data mining problems, scalability with respect to data and size is an important new criterion. It indicates innovative and totally new approaches to information management. It is important for all organizations that utilized large data sets. Any organization with large volumes of databases or helpdesk service records can benefit from this newly emerging concept. It is the actual discovery phase of knowledge discovery process.

1. Scope of Data Mining

Data mining derives its name from the similarities between searching for valuable information in a large database. In the given databases of sufficient size and quality, data mining technology can generate new business opportunities by providing these capabilities:

· **Automatic prediction of trends and behaviors:** It automates the process of finding predictive information in large databases. Questions that traditionally required extensive hands-on analysis can now be answered directly from the data - quickly. A typical example of a predictive problem is targeted marketing. Data mining uses data on past promotional mailings to identify the targets most

likely to maximize return on investment in future mailings. Other predictive problems include forecasting bankruptcy and other forms of default, and identifying segments of a population likely to respond similarly to given events.

· **Automatic discovery of previously unknown patterns:** Data mining tools sweep through databases and identify previously hidden patterns in one step. An example of pattern discovery is the analysis of retail sales data to identify seemingly unrelated products that are often purchased together. Other pattern discovery problems include detecting fraudulent credit card transactions and identifying *anomalous data* that could represent data entry keying errors.

Data mining techniques can yield the benefits of automation when implemented on existing software and hardware platforms, and can be implemented on new systems as existing platforms are upgraded and new products developed. When data mining tools are implemented on high performance parallel processing systems, they can analyze massive databases in minutes. Faster processing means that users can automatically experiment with more *models* to understand complex data. High speed makes it practical for users to analyze huge quantities of data. Larger databases, in turn, yield improved predictions. Databases can be larger in two senses:

· **Higher dimensionality :** In hands-on analyses, analysts must often limit the number of variables they examine because of time constraints. Yet variables that are discarded because they seem unimportant, may carry information about unknown patterns. High performance data mining allows users to explore the full dimensionality of a database, without preselecting a subset of variables.

· **Larger samples:** Larger samples yield lower estimation errors and variance, and allow users to make inferences about small segments of a population.

Gartner Group Advanced Technology Research Note listed data mining and Artificial Intelligence at the top of the five key technology areas that "will clearly have a major impact across a wide range of industries within the next three to five years." Gartner also listed parallel architectures and data mining as two of the top ten new technologies in which companies will invest during the next five years.

2. Data Mining Working

Data mining can be distinguished from other retrieval technologies in that it makes choices and calculations for the searcher and then categorizes information based on those choices. It accomplishes this by identifying data relevant to meet users' information needs, and then organizing documents by topic, source, relationship with other documents, and a number of other criteria.

i. The first step that any data mining tool must accomplish is to identify which documents should be searched. In some cases, a known body of documents such as a magazine or image database may be searched. In other cases such as in the world wide web, unfamiliar documents and services will be searched. The determination of which documents to search depends on knowledge of what the users intend to do with the information they find.

ii. Once the data mining software has determined which documents it should search, it must then extract and normalize data that are relevant to the query. For text documents, stemming algorithms, grammar parsers, idiom detectors, thesauri, or other methods might be applied on the search terms as well as the documents searched to ensure results that are more relevant and comprehensive than could be accomplished by string or regular expression matching. It is at this step that data are categorized for use by the data-mining algorithm. This step is roughly analogous to automatic authority control in a library setting.

iii. After the data is prepared, the algorithms that search and arrange the data must be determined. The choice of the data-mining algorithm depends at least partly on the purpose for the search. For example, if user types in a personal name, the data-mining algorithm might separate the output into categories such as biographical information, graphical files (i.e., pictures of the person), and documents authored by the person.

Data mining is not so much a single technique as the idea that there is more knowledge hidden in data than show itself on the surface. So it is really an `anything goes' affair. So any technique that helps in extracting more out of the data is useful hence data mining is formed of a group of heterogeneous tasks. Data mining algorithms vary from organization to organization but most commonly used techniques in data mining are:

· **Artificial neural networks:** These are non-linear predictive models that learn through training and resemble biological neural networks in structure.

· **Genetic algorithms:** These are 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:** These are 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 neighbour method:** This is 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 neighbour technique.

· **Rule induction:** The rule induction is the extraction of useful if-then rules from data based on statistical significance.

· **Data visualization:** This is the visual interpretation of complex relationships in multidimensional data.

In a library setting, one or more of the following patterns are to be followed:

· **Classification and Clustering:** Classification mimics library cataloging procedures by grouping structured and unstructured data according to certain criteria such as source (e.g., government bodies), document type (e.g., maps), language, subject, or a number of other criteria. Clustering is similar to classification, except that the classes are determined by finding natural groupings in the data items based on probability analyses rather than by predetermined groupings. Clustering and classification are often used as a starting point for exploring further relationships in data. For example, many Internet search engines (such as Northern Light) break down sites by location, subject, or language before sub-arranging data.

· **Link Analysis:** Like wise the paper materials, where similar documents tend to have similar bibliographical references, and frequency of citation is often considered to reflect the quality or importance of a document, link analysis assumes that higher-quality or otherwise more-desirable documents will generally be linked to more frequently than other documents, and that links in a document reveal something about the content of a document. Link analysis can place frequently linked-to documents at the top of a list or identify documents that are associated with each other.

· **Sequence Analysis:** Sequence analysis uses statistical analysis to identify unlinked documents that users are likely to want to read together. It examines the paths that users follow when searching for information and can help identify which documents users are likely to want together.

· **Summarization:** Though machine-generated abstracts are inferior to human-generated ones in terms of readability and content, yet they can be very useful for helping users decide what items they need. Abstract-generating software typically works by identifying significant words or phrases based on position within documents, association with critical phrases.

3. Data Mining and the Libraries

Though 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 and its use in libraries is limited. Data mining offers two major advantages to libraries:

(1) Firstly, it can provide faster and more thorough access to materials than that provided by manual cataloging; and

(2) Secondly, it can be used by employees or users with basic computer and analytical skills, so people can more easily find what they need without the assistance of highly skilled staff. Data mining also has some drawbacks as data mining tools are not standardized and vary greatly in effectiveness. Besides, the technology is largely untested in a library setting. Most of successful projects involved statistical data or relatively short records not the lengthy text documents and multimedia objects from a variety of sources that library users frequently seek.

4. Problems with Data Mining

There are some problems with data mining technology, of which main are:

· **Lack of Standards:** The most serious problem is that there are no established standards for data mining storage and retrieval. Besides, the record sharing between libraries is impractical, and long-term access to materials is in doubt. In an electronic environment where database access in a library is determined by IP or password validation, record sharing may not be an important consideration, but information-sharing programs such as interlibrary loan become complicated at best and impossible at worst. Because data mining increases library dependence on proprietary functions, libraries that invest heavily in data mining technologies increase the risk of incurring expensive and difficult conversions or severe data loss when vendors quit supporting their products. In the present environment, world wide use of the MARC format dramatically has reduced data migration problems and greatly simplified record sharing and interlibrary loan.

· **Unproven to Libraries:** It is unclear whether data mining techniques used on the Internet or for certain business and scientific applications can be successfully applied in a library setting. In contrast to data mining in the business and scientific communities involve short documents consisting of well-structured or statistically oriented data, libraries work predominantly with large unstructured text documents from diverse sources. While a number of text-mining tools do provide access to minimally structured text documents, the total amount of information they provide access to is small in comparison with that found in a large library. Also, Web pages, e-mail, and corporate reports (the focus of most text-mining tools) tend to be relatively short, so the procedures used to index and search them might not work successfully with the larger information objects typically found in libraries.

· **Technical Hurdles:** The other problem with data mining is that it faces the same difficulties as other searching mechanisms. The quality of data is critical for successful data mining, just as it is for successful searching by other methods. If information is not structured in a way that allows pattern discovery, the likelihood of extracting meaningful information from the data is greatly reduced. Data mining looks for patterns in data. It is very difficult for data mining tools to identify the relationships between different information objects when it is not possible to determine the meaning of the data. Despite of advancement in technology, it is not practical to use all processing techniques on all documents in a given search, except when small sets of data are concerned. Unless all data can be stored in memory and there is sufficient processing power, heuristics must be used to determine the optimal searching strategy. Users may reveal information about themselves and the purpose of their searches in the way they phrase their queries, but it is difficult to glean enough information to identify techniques that will optimally serve the user.

Moreover, effective techniques for indexing and retrieving non-textual data are not yet available. As the number of multimedia information objects increases rapidly, so will the need for effective storage and retrieval mechanisms. When this problem is considered together with the lack of storage and retrieval standards even for text documents, libraries need to be wary of depending on particular data mining technologies that are not expected to provide long-term access to materials.

· **Inappropriateness of Data-mining Tools:** Before committing to data mining technologies on a large scale, libraries need to determine how data mining fits with existing resources and organizational goals. Generally speaking, data mining technologies are most beneficial to libraries that are interested in purchasing access to databases rather than physical materials. Full-text, dynamically changing databases tend to be better suited to data mining technologies than the online catalog, which is cumbersome and expensive to update. On the other hand, libraries concerned with providing long-term access to physical items that exist within the library would be well advised to adopt a sit-and-wait attitude at this point especially since good access to these materials is provided through the online catalogue.

5. CONCLUSION

Practically, data mining tools have neither proven effective at integrating data from different sources nor have they proven effective with non-textual data nor have they found new ways to present relationships between information in large retrieval sets that make sense to users beyond a primitive level. Moreover, patrons and staff alike can get confused if a library gets involved with a wide variety of storage and retrieval mechanisms. There is a potential lack of long-term access and an inability to share certain resources with other libraries problematic and lack of indexing and retrieval standards puts long-term access to materials in doubt and severely undermines the ability of the library to share its resources. Inspite of above constraints, data mining concept alike business field, is getting attention from the library experts and definitely, libraries will be benefited with this technique in future.

REFERENCE

1) Adriaans, P. and Zantinge, D. 2001. Data Mining. Pearson Education Asia, Delhi.

2) Anahory, S. and Murray, D. 2001. Data Warehousing in the Real World: A Practical Guide for Building Decision Support Systems. Pearson Education Asia, Delhi.

3) An Overview of Data Mining at Dun & Bradstreet. DIG White Paper 95/01. September 1995. Data Intelligence Group, Pilot Software.

4) AUUG: Annual Winter Conference '97 Data Mining on the World Wide Web Enno Davids, Metva P/L. Enno.Davids@metva.com.au mailto:Enno.Davids@metva.com.au.

5) Banerjee, K. 1998. Is Data Mining Right for Your Library? *Computer in Libraries.* Vol. 18 No 10 (November/December).

6) Dhiman, A.K. 2002. Knowledge Management System for Knowledge Management in Information Technology Era. *Indian Journal of Information, Library & Society*. 15: (communicated).

7) Dhiman, A.K. 2003. Basics of Information Technology for Librarians and Information Scientists. Vol. 2nd. Ess Ess Publications, Delhi. pp. 262-68.

8) Grimshaw, D.J. 2000. Bringing Geographical Information Systems into Business. 2nd Edition. John Wiley & Sons Inc., New York. p. 212.

9) Srikant, R. Data Mining Technologies for Digital Libraries and Web Information Systems. IBM Almaden Research Center. 650 Harry Road. San Jose, CA 95120, USA. srikant@us.ibm.com.

10) Tiwana, A. 2000. The Knowledge Management Toolkit. Pearson Education Asia, Delhi. P. 216.

BRIEF BIOGRAPHY OF AUTHOR

Dr. Anil Kumar Dhiman holds M.A., M.Sc., MLISc., B.Ed., P.G.D.C.A. and Ph.D. Degree in Botany. He is Fellow and Life member of various professional associations and has over four dozens papers and nine books to his credits in the both field of his study, Library & Information Science and Botany. At present, he is working in the Central Library of Gurukul Kangri Univeristy, Hardwar. He has also been awarded with APSI Young Scientist Award and Gold Medal in 1999.