Use of Data Minining in a Library

Rakesh Kumar Mishra

Programmer-Cum-System Incharge (Information Scientist)

Central Library, D.D.U.Gorakhpur University, Gorakhpur-273009, U.P.

Abstract

Data mining is a form of artificial intelligence(AI) that uses automated process to finding interesting information in large repositories of data. The term data mining also refers to the step in the knowledge discovery process in which special algorithms are employed in hopes of identifying interesting patterns in the data. These interesting patterns are then analyzed yielding knowledge. Any information system design for the libraries with the help of data mining is most useful for librarian's to make appropriate decision . The desired outcome of data mining activities is to discover knowledge that is not explicit in the data, and to put that knowledge to use. Libraries are already benefiting from data mining techniques as they explore ways to automatically classify information and explore new approaches for subject clustering. As the field grows, new applications for libraries are likely to evolve and it will be important for library administrators to have a basic understanding of the technology. This paper is mainly written for the use of data mining to management decision in libraries. It explains the future development should focus on developing tools and techniques that yield useful knowledge without invading individual privacy.

Introduction

Data mining is a tool and some times called as a discovery-oriented data analysis (DODA) technology and not a single product or a system. It itself is in the 2nd generation of AI by which libraries will continue to satisfy user's 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.

Data mining refers to "using a variety of techniques to identify nuggets of information or decision-making knowledge in bodies of data, and extracting these in such a way that they can be put to use in the areas such as decision support, prediction, forecasting and estimation. The data is often voluminous, but as it stands of low value as no direct use can be made of it; it is the hidden information in the data that is useful"

The following diagram summarises the some of the stages/processes identified in data mining and knowledge discovery by Usama Fayyad & Evangelos Simoudis, two of leading exponents of this area.

This phases depicted start with the raw data and finish with the extracted knowledge which was acquired as a result of the following stages:

- Selection - selecting or segmenting the data according to some criteria e.g. all those people who own a car, in this way subsets of the data can be determined.
- Preprocessing - this is the data cleansing stage where certain information is removed which is deemed unnecessary and may slow down queries for example unnecessary documents which are not useful for circulation purpose.
- Transformation - the data is not merely transferred across but transformed in that overlays may added such as the demographic overlays commonly used in market research. The data is made useable and navigable.

Reprint requests: Rakesh Kumar Mishra
Programmer-Cum-System Incharge
(Information Scientist), Central Library D.D.U.
Gorakhpur University, Gorakhpur-273009 U.P.)
Email : rkmishra42@yahoo.com

- Data mining - this stage is concerned with the extraction of patterns from the data. A pattern can be defined as given a set of facts(data) F, a language L, and some measure of certainty C a pattern is a statement S in L that describes relationships among a subset Fs of F with a certainty c such that S is simpler in some sense than the enumeration of all the facts in Fs.
- Interpretation and evaluation - the patterns identified by the system are interpreted into knowledge which can then be used to support human decision-making e.g. prediction and classification tasks, summarizing the contents of a database or explaining observed phenomena.

Data mining is a form of artificial intelligence which uses automated processes to find information that users want. Although its use in libraries is limited, 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.

Data mining offers two potential advantages to libraries firstly, it can provide faster and more thorough access to materials than that provided by manual cataloging; and 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. However, data mining also has drawbacks. Data mining tools are not standardized and vary greatly in effectiveness. Also, the technology is largely untested in a library setting. Most 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.

1.1 How Exactly Data Mining Work?

Before explaining how the data mining work exactly, it is better to understand the concept of various steps for extracting hidden knowledge from the data of a particular library or organization, which are as follows-

(1). Identify the objective ⟶ (2). Select the data ⟶ (3). Prepare the data for use ⟶ (4). Evaluate the data in order to choose the appropriate tools ⟶ (5). Format the Solution ⟶ (6). Select the tools ⟶ (7). Construct the model ⟶ (8). Validate the Findings ⟶ (9). Deliver the Findings ⟶ (10). Integrate the solutions

These above various steps use not only for library but also for all organization to dig the data for extracting meaningful information to make appropriate decision for the library or any organization.

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 a user's information needs and then organizes documents by topic, source, relationship with other documents and a number of other criteria.

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 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 user intends to do with the information s/he finds.

For example, computers can be programmed to recognize personal and place names as well as to which part of speech a search term belongs. When a user is seeking information about a person, it is reasonable for data mining software to search for images of the person. Likewise, if the object of the search is a place, it is logical for data mining software to search for a map, though it would make little sense to search for through images of people. While it is often not possible to make assumptions about users' goals determine, users often convey information about themselves and their needs in the queries which can be used by data mining tools.

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.

After the data are prepared, the algorithms which 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 a 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 algorithms vary, but in a library setting, these algorithms are likely to follow one or more of the following patterns : **A.Classification and Clustering**

Classification problems aims to identify the characteristics that indicate the group to which each case belongs and also it mimics library cataloging procedures by grouping structured and unstructured data according to certain criteria such as source, document type , 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 analysis rather than by predetermined groupings. Clustering differs from classification in that it does not rely on predefined classes or characteristics for each group. 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 subarranging data.

B. Link Analysis

When paper materials are concerned, similar documents tend to have similar bibliographical references and frequency of citation is often considered to reflect the quality or importance of a document. Similarly, 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.

C. Association/Sequence Analysis

Sequence analysis uses statistical analysis to identify unlinked documents that users are likely to want to read together. To illustrate this principle, consider the organization of a discount store. Shampoo is typically found closer to other hair care products such as combs than it is to chemically similar products such as dish detergent. Sequence analysis examines the paths that users follow when searching for information and can help identify which documents users are likely to want together.

D. Summarization

Even though machine generated abstracts are inferior to human generated ones in terms of readability and content, 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 document, association with critical phrases, syntactical analysis, grammar parsing, and other methods.

The most commonly used techniques in Data Mining are :

1. Artificial neural networks : Non-linear predictive models that learn through training and resemble biological neural networks in structure.
2. Decision trees: Tree-shaped structures that represent sets of decision, which generates rules for the classification of a dataset. CART (Classification & Regression Trees) & CHAID(Chi Square Automatic Interaction Detection) are important specific decision tree methods.
3. Genetic Algorithm : Optimization techniques that use processes such as genetic combination, mutation, and natural selection in a design based on the concepts of evolution.
4. Nearest neighbor method : It is a technique that classifies each record in a dataset based

on a combination of the classes of the k records(s) most similar to it in a historical dataset.

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

An Example of Data Mining

Although Internet search engines typically don't describe themselves as data mining tools, they often perform data mining in order to ensure that search results are far superior to results that could be generated by keyword or even nested Boolean searching. This Northern Light retrieval set appears easy to navigate even though over 50 thousand documents were found. The "custom search folders" group documents according to source, subject matter, and type of document and allow the searcher to locate the information s/he seeks.

In this particular example, a person wishing to understand how electronic course reserves work would be well advised to check in the "questions and answers" folder but administrators looking to buy electronic course reserves systems could probably search within the "commercial sites" folder. Within each category, documents are further subcategorized so that a searcher can reasonably browse through large retrieval sets to find a specific document. Based on the search output, it is also clear that the search engine realized that the search term was an idiom so that unrelated documents that just happened to contain the search terms could be ranked much lower in the list or separated out by subject.

Data mining tools differ significantly in terms of purpose and functionality. On the Internet, there are many kinds of data, but the major Internet search engines can only search text files - they cannot read documents in other formats. As such, search engines represent only one kind of text processing tools. Most tools are oriented towards business purposes such as processing marketing data or web transaction logs. Virtually all tools are limited to processing alphanumeric data - good tools for processing multimedia data are not yet available.

It is important to note that data mining is not a linear process because associations can occur at multiple levels of abstraction so various processing steps may need to be repeated. To imagine what this might look like, consider a search like the Northern Light search above. If the results are presented in folders according to site, subject matter, and popularity, it might have been necessary to use classification, clustering and link analysis simultaneously. Also, it is important to keep in mind that a particular document might belong in more than one place within the set of retrieved documents.

1.2 Data mining problems/issues

Data mining systems rely on databases to supply the raw data for input and this raises problems in that databases tend be dynamic, incomplete, noisy, and large. Other problems arise as a result of the adequacy and relevance of the information stored, which are as follows-

A. Limited Information

A database is often designed for purposes different from data mining and sometimes the properties or attributes that would simplify the learning task are not present nor can they be requested from the real world. Inconclusive data causes problems because if some attributes essential to knowledge about the application domain are not present in the data it may be impossible to discover significant knowledge about a given domain.

B. Noise and missing values

Databases are usually contaminated by errors so it cannot be assumed that the data they contain is entirely correct. Attributes which rely on subjective or measurement judgements can give rise to errors such that some examples may even be misclassified. Error in either the values of attributes or class information are known as noise. Obviously where possible it is desirable to eliminate noise from the classification information as this affects the overall accuracy of the generated rules.

C. Uncertainty

Uncertainty refers to the severity of the error and the degree of noise in the data. Data precision defined as the ratio of relevant items out of total retrieved items, is an important consideration in a discovery system.

D. Size, updates, and irrelevant fields

Databases tend to be large and dynamic in that their contents are ever-changing as information is added, modified or removed. The problem with this from the data mining perspective is how to ensure that the rules are up-to-date and consistent with the most current information. Also the learning system has to be time-sensitive as some data values vary over time and the discovery system is affected by the 'timeliness' of the data.

D. Lack of Standards

Significant obstacles must be overcome to implement data mining in a library setting. The most serious of these is probably that there are no established standards for data mining storage and retrieval. Consequently, record sharing between libraries is impractical and long term access to materials is in doubt. 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, widespread use of the MARC format dramatically reduces data migration problems and greatly simplifies record sharing and interlibrary loan.

E. It's Unproven in Libraries

It is unclear that data mining techniques used on the Internet or for certain business and scientific applications can be successfully applied in a library setting. Most successful examples of data mining in the business and scientific communities involve short documents consisting of well-structured or statistically-oriented data. Conversely, libraries work predominantly with large unstructured text documents from diverse sources. While a number of text mining tools do provide access 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 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.

F. Big Technical Hurdles Remain

The last 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 advances 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 their 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 queries, but it is difficult to glean enough information to identify techniques which 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.

1.3.Whether Data Mining Tools are Appropriate in the Library?

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 which 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 catalog.

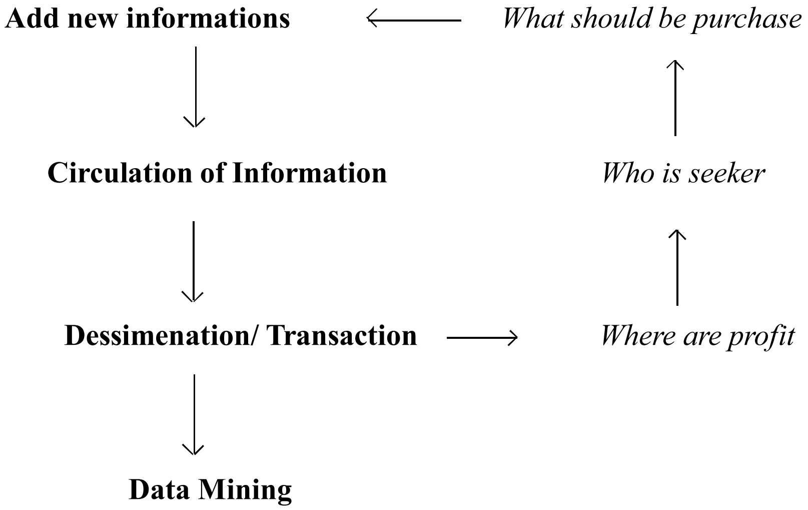
Before investing in data mining technology, a library needs to ask itself a number of practical questions. For example, what does the library need to do with the data before it can be searched? Does the tool provide easy to understand comprehensive search results for users? Data mining tools have neither proved effective at integrating data from different sources nor have they proved effective 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. Is a potential lack of long term access and an inability to share certain resources with other libraries problematic? Lack of indexing and retrieval standards put long term access to materials in doubt and severely undermine the ability of the library to share its resources.

A great number of initiatives have been taken to implement data mining, but the only areas that have been successful so far are those which contain well-defined data and modest goals. Data mining technologies are still in a very early stage of development. While data mining has been used very successfully in the business world for statistically oriented tasks such directed advertising, progress towards the same success with descriptive data has been slower.

Consider the example of library system which enhance by data mining techniques are

1. Increase the demand of relevant information.
2. Attract more information seeker to get his/ her require information.
3. Library & Information science professionals on making the policy to provide right information to right user at right time in right form.

In this context the relationship between Data mining and Library Management can be shown by following figure-



Libraries should continue to explore alternatives to the traditional catalog for providing access to materials. Automated indexing systems can improve access to materials by continuously updating outdated information. Even though print resources will remain an important component of the library's collection, an increasing number of resources will be available in electronic form only. Traditional access and indexing mechanisms based upon a print-based world do not provide sufficient access to the diversity of data types and structures that exist in a digital environment. Furthermore, there is simply too much information to process manually, so increased reliance on automated access mechanisms such as data mining tools seems to be just a matter of time.

Conclusion

Data mining offers great promise in helping organizations uncover patterns hidden in their data that can be used to predict the behaviour of user.Data mining & knowledge discovery are both synonym word, which refers to a specific step in the knowledge discovery process, that focuses on the application of specific algorithms, is the mechanism to create mining models used to identify interesting patterns for effective management decision in libraries . These patterns are then conveyed to user who converts these patterns into useful knowledge and utilize that knowledge.

References

1. Two Crows : Data mining Glossary.
2. W.H. Inmon. "The Data Warehouse And Data Mining." Communications of the ACM, Nov. 1996, v. 39, no. 11, p. 49.
3. Dunham, M.H. (2003). Data mining introductory and advanced topics
4. Han, J., & Kamber, M. (2001). Data mining: concepts and techniques
5. Deborah Asbrand. " Is Datamining Ready For The Masses?"