Use of Data Mining in the field of Library and Information Science : An Overview

Roopesh K Dwivedi R P Bajpai

Abstract

Data Mining refers to the extraction or "Mining" knowledge from large amount of data or Data Warehouse. To do this extraction data mining combines artificial intelligence, statistical analysis and database management systems to attempt to pull knowledge form stored data. This paper gives an overview of this new emerging technology which provides a road map to the next generation of library. And at the end it is explored that how data mining can be effectively and efficiently used in the field of library and information science and its direct and indirect impact on library administration and services.

Keywords : Data Mining, Data Warehouse, OLAP, KDD, e-Library

0. Introduction

An area of research that has seen a recent surge in commercial development is data mining, or knowledge discovery in databases (KDD). Knowledge discovery has been defined as "the non-trivial extraction of implicit, previously unknown, and potentially useful information from data" [1]. To do this extraction data mining combines many different technologies. In addition to artificial intelligence, statistics, and database management system, technologies include data warehousing and on-line analytical processing (OLAP), human computer interaction and data visualization; machine learning (especially inductive learning techniques), knowledge representation, pattern recognition, and intelligent agents.

One may distinguish between data and knowledge by defining data as corresponding to real world observations, being dynamic and quite detailed, whereas knowledge is less precise, is more static and deals with generalizations or abstraction of the data [2]. A number of terms have been used in place of data mining, including information harvesting, data archaeology, knowledge mining, and knowledge extraction. The knowledge is stored in data warehouse, which is the central store house 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 historic in nature, so they contain extremely large datasets [3].

Libraries also have the big collection of information and in e-Library there are organize collection of information which serves a rich resource for its user communities. E-Library includes all the processes and services offered by traditional libraries though these processes will have to be revised to accommodate difference between digital and paper media. Today's e-Libraries are built around Internet and Web technologies with electronic books and journals as their basic building blocks. Here Internet serves as a carrier and provides the contents delivery mechanism and Web technology provides the tools and techniques for content publishing, hosting and accessing. The availability of computing power that allow parallel processing, multitasking and parallel knowledge navigation with increasing popularity of Internet and development in Web technologies are the main catalyst to the concept of e-Library.

Data Mining is relatively new term in the world of library and information science though it is being used by both commercial and scientific communities since a long time. There are three main reasons for that. First both the number and size of databases in many organizations are growing at a staggering rate. Terabyte and even petabyte databases, once unthinkable, are now becoming a reality in a variety of

domains, including marketing, sales, finance, healthcare, earth science, molecular biology (e.g. the human genome project), and various government applications. Second organizations have realized that there is valuable knowledge which is buried in the data which, if discovered, could provide those organizations with competitive advantage. Third, some of the enabling technologies have only recently become mature enough to make data mining possible on large datasets.

1. Data Mining: The Concept

"Data mining is the exploration and analysis, by automatic and semiautomatic means, of large quantities of data in order to discover meaningful patterns and rules"[5]. It takes data and business opportunities and produce actionable results.

1.1 The Knowledge Discovery Database (KDD) Process

The data mining is actually a step in a larger KDD process. The KDD process employs data mining methods or algorithms to extract or identify knowledge according to some criteria or measure of interestingness, but it also includes steps that prepare the data, such as preprocessing, sub-sampling, and transformations of the database [6].

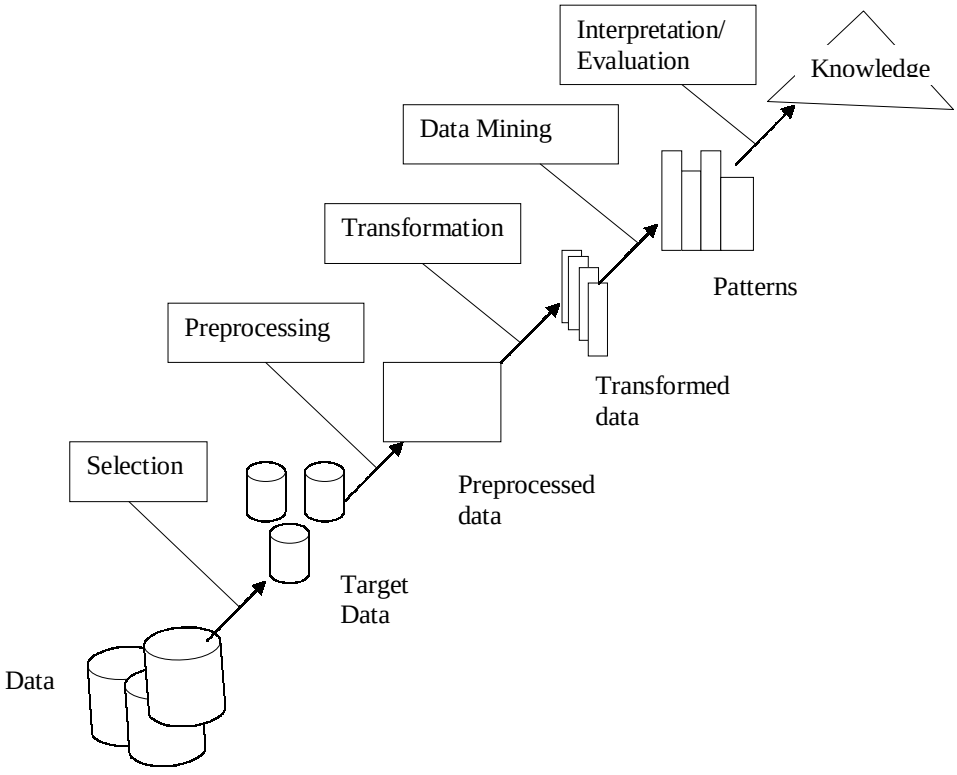

Figure1. The KDD Process

The first step in the KDD process is to select data to be analyzed from the set of all available data. In many cases, the data is stored in transaction databases. These databases are quite large and extremely dynamic. Therefore a subset of the data must be selected from those databases, since it is unnecessary in the early stages to attempt to analyze all data.

Target data is then moved to a cache or another database for further preprocessing. Preprocessing is extremely important step in KDD process. Often, data have errors introduced during the input process, either from a data entry clerk entering data incorrectly or from a faulty data collection device. If target data are being extracted from several source databases, the databases can often be inconsistent with each other in terms of their data models, the semantics of the attributes, or in the way the data is represented in the database. If the two databases were built at different times and following different guidelines, it is entirely possible that they may be two different data models (relational and object-oriented) and two different representations of the entities or objects and there relationships to each other. The preprocessing step should identify these differences and make the data consistent and clean.

The data can often be transformed for use with different analysis techniques. A number of separate tables can joined into one table, and vice versa. An attribute that may be represented in two different forms (date written as 3/15/97 versus 15-3-1997) should be transformed into common format. If the data is represented as text, but it is intended to use a data mining technique that requires the data to be in numerical form, the data must be transformed accordingly.

At this point, data mining algorithms can be used to discover knowledge, e.g., trends, patterns, characteristics, or anomalies. The appropriate discovery or data mining algorithms should be identified, as they should be pertinent to the purpose of the analysis and to the type of data to be analyzed. Often, the data mining algorithms work more effectively if they have some amount of domain information available containing information on attributes that have higher priority than others, attributes that are not important at all, or established relationships that are already known. Domain information is often collected in knowledge base, a storage mechanism similar to a database but used to store domain information and other knowledge.

When a pattern is identified, it should be examined to determine whether it is new, relevant and “correct” by some standard of measure. The interpretation and evaluation step may involve more interaction with a user or with some agent of the user who can make relevancy determinations. When the pattern is deemed relevant and useful, it can be deemed knowledge. The knowledge should be placed in the knowledge base for use in subsequent iterations. Note that the entire KDD process is iterative; at many of the steps, there may be need to go back to a previous step, since no patterns may be discovered, new data should be selected for additional analysis, or the patterns that are discovered may not be relevant.

In many step of KDD process, it is essential to provide good visualization support to the user. This is important for two reasons. First , without such visualizations, it may be difficult for users to determine the usefulness of discovered knowledge-often a picture is worth a thousand words. Second, given good visualization tools, the user can discover things that automated data mining tools may be unable to discover. Working as a team, the user and automated discovery tools provide far more powerful data mining capabilities than either can provide alone.

1.2 The Virtuous Cycle of Data Mining

The promise of data mining is to find the interesting patterns in the large amount of database. But merely finding the patterns is not enough. You must be able to respond to the patterns, to act on them, ultimately turning the data into information, the information into action, and the action into the value. This is the virtuous cycle of data mining in a nutshell.

There are four stages of the virtuous cycle of data mining.

i. Identify the business problem
ii. Use data mining technique to transform the data into actionable information
iii. Act on the information
iv. Measures the result

These steps are highly interdependent; the results of one stage are the inputs into the next phase, much like the steps in multi-step manufacturing process. The whole approach is driven by results. Each stage depends on the results from the previous stage.

2. Use of Data Mining

Data mining derives its name from the similarities between searching for valuable information in a large database. Granter 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 with in the next three to five years." Granter 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 [3].

In fact practical data mining can accomplish a limited set of tasks and only limited circumstances. The main important thing is that it can be used in many problems of intellectual, economic, and business interest. These problems can be phrased in terms of the following six tasks.

i. Classification- classification consists of examining the features of newly presented object and assigning it to one of the predefined set of classes.
ii. Estimation- classification deals with discrete outcomes while estimation deals with continuously valued outcomes. Given some input data, we use estimation to come up with a value for some unknown continuous variable such as income, height, or credit card balance.
iii. Prediction- prediction is the same as classification or estimation except that the records are classified according to some predicted future behavior or estimated future value. In prediction task, the only way to check the accuracy of the classification is to wait and see.
iv. Affinity Grouping- the task of affinity grouping is to determine which things go together. Affinity grouping can also be used to identify cross-selling opportunities and design attractive packages or grouping of product and services. Affinity grouping is one simple approach to generating rules from data. If two items, say cat food and kitty litter, occur together frequently enough, we can generate two association rules:

    - People who by cat food also by kitty litter with probability P1.
    - People who by kitty litter also by cat food with probability P2.

v. Clustering- clustering is the task of segmenting a heterogeneous population into a number of more homogeneous subgroups or clusters. Clustering differs from classification in the way that clustering does not rely on predefine classes. In clustering, there are no predefined classes and no examples. The records are group together on the basis of self-similarity.
vi. Description- some time the purpose of data mining is simply to describe what is going on in complicated database in a way that increases our understanding of the people, products, or processes that produced the data in the first place. A good enough description of a behavior will often suggest an explanation for it as well.

No single data mining tool and technique is equally applicable to all the tasks. In commercial application, data mining is usually employed on very large databases. The reasons for this are two fold.

- In small databases, it is possible to find interesting patterns and relationships by simple inspection of results from familiar tool such as spreadsheets and multidimensional query tools.
- Most data mining technique require large amount of training data containing many examples in order to generate classification rules, association rules, clusters, or predictions. Small datasets lead to unreliable conclusion based on chance patters.

3. Data Mining Methodology

Data mining process can be divided into four stages:

i. Identify the problem
ii. Analyzing the data
iii. Taking action
iv. Measuring the outcome

The first and third stages raise mainly business issues. For data mining to be successful, these business issues must, of course properly addressed.

There are two basic style of data mining

- Hypothesis testing- is a top down approach that attempts to substantiate or disprove preconceived ideas.
- Knowledge Discovery- is a bottom approach that starts with the data and tries to get it to tell us something we didn't know.

3.1 Hypothesis Testing

Hypothesis testing is what scientists and statisticians spend their lives doing. Testing the validity of a hypothesis is done by analyzing data that may simply be collected by observation or generated through an experiment, such as test mailing. The hypothesis testing method has several steps:

i. Generate good ideas (hypothesis)
ii. Determine what data would allow these hypotheses to be tested.
iii. Locate the data
iv. Prepare the data for analysis
v. Build computer model based on the data
vi. Evaluates computers models to confirm or reject hypothesis

3.2 Knowledge Discovery

Knowledge discovery can be either directed or undirected. Directed Knowledge Discovery is goal oriented. There is a specific field whose value we want to predict, a fix set of classes to be assigned each record, or a specific relationship we want to explore. The directed knowledge discovery method has several steps :

i. Identify sources of pre-classified data
ii. Prepare data for analysis
iii. Build and train the computer model
iv. Evaluate the computer model
v. Apply the Computer model to the new data.

In Undirected Knowledge Discovery the data mining tool is simply let loose on the data in the hope that it will discover meaningful structure. One common use of undirected knowledge discovery is market basket analysis. Another application is clustering, where groups of records are assigned to the same cluster if they have some thing in common. The undirected knowledge discovery method has several steps :

i. Identify source of data
ii. Prepare data for analysis
iii. Build and train a computer model
iv. Evaluate the computer model
v. Apply the computer model to the new data
vi. Identify potential targets for directed knowledge discovery
vii. Generate new hypothesis to test.

You will notice that steps I through V are the same as for directed knowledge discovery. The two additional steps reflect the fact that undirected knowledge discovery is usually a prelude to further investigation via more directed technique.

4. Data Mining and the Libraries

Till now we have discusses about the data mining and its working. Now we are going to explore how data mining can be useful in the field of library and information science. As per fifth law of library science “Library is a growing organization” [9] so the volume of the library data is also growing with an enormous rate. For efficiently and effectively doing the library administration and extending library services the need of library automation and e-Library occur. But simply automating the library or developing an e-Library is not the only solution unless and until we are not able to explore the hidden information from the large amount of database. This can be done by applying the data mining in the library data.

Now we take a glance on the possibilities opening in the new age of data mining in the field of library and information science.

i. Classification - By using data mining we can develop a computer program that will replace the manual classification with the automatic classification of library contents. 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 [3].
ii. 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 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 ac 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 [3].

iii. 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 [3].

iv. 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].

v. Clustering- 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 search engine (such as Northern Light) break down sites by location, subject, or language before sub-arranging data [3].

5. Future of data mining in the library working

In future Data Mining can provide the new road map for the next generation of library by applying it for the following activities of library.

i. Searching of Information (Reference Service)- Since the data of the library continuously growing with an exponential rate and the main problem is how one can reference the required information form the large amount of redundant information of the library. This can be possible by applying data mining techniques, so one can say that the data mining is the future of reference service.

ii. Classification- It will replace the manual classification of content of the library with the computer assisted classification, so that the classification task can be accomplished by less skilled person in a fast and efficient way. This will simplify the classification task of the library.

iii. Acquisition- As per third law of library science “Every book its reader” [9]. By applying the data mining in the library data it can be easily find out the required contents that are necessary to acquire next. This will reduce the work of library staff related to the acquisition as well as the efficient use of budget allocated to the library.

6. Conclusion

It can be concluded that there is the need of data mining techniques that will redesign and simplify the working of library like classification, acquisition, circulation and referencing. The main use of data mining is in referencing but it can be used for some other work of library as well. So it is urgently needed that systematic efforts have been take place to develop data mining techniques and algorithms for library database.

7. References

1. Frawley, W., Piatetsky-Shapiro, G., Matheus, C. J., “Knowledge Discovery in Database an Overview,” in knowledge Discovery in Databases, G. Piatetsky-Shapiro and W. Frawley, (Eds.), MIT Press, 1991
2. Wiederhold, G., “Knowledge Versus Data,” in on Knowledge Base Management Systems: Integration Artificial Intelligence and Database Technologies, Brodie, M. and Mylopoulos, J. (Eds.), Springer-Verlag, 1986.
3. Dhiman, Anil K., Data Mining and its use in Libraries, CALIBER-2003.
4. Carbone, Patricia L., Data Mining or “Knowledge Discovery in Databases” : An Overview, MITRE Corporation, 1997

5. Berry, Michael J.A., Linoff, G., Data Mining Techniques for Marketing, Sales and Customer Support, Wiley Computer Publishing.
6. Fayyed, U., R. Uthursamy, (Eds.), Proceeding of the first international conference on knowledge discovery and data mining, The AAAI Press, Menlo Park, CA, 1995.
7. Han, J., Kamber, M., Data Mining: Concept and Techniques, Morgan Kaufmann Publishers, 2002.
8. Adriaans, P., Zantinge, D., Data Mining, Pearson Education, 2003.
9. Rangnathan, S.R., Five laws of library science, Sarda Rangnathan Endowment for Library Science, Banglore 1993.

About Authors

Mr. Roopesh K. Dwivedi is Information Scientist at MGCGV, Chitrakoot, Satna, India.
E-mail: rkdwivedi_mgcgv@rediffmail.com

Mr. R.P. Bajpai is In-charge Librarian at MGCGV, Chitrakoot, Satna, India.
E-mail: rpbajpai_mgcgv@rediffmail.com