Data Mining in the Process of Knowledge Discovery in Digital Libraries

Mahesh Mudhol Purushothama Gowda

Abstract

Data Mining is the partially automated process of extracting patterns, usually from large data sets. In this paper the authors tries to give an overview about data mining.

Keywords : Data Mining, Digital Library, Knowledge Management.

0. Introduction

A digital library is an advanced library system that stores information in the digital form and that provides facilities to search and access information. A core part of the digital library is the DBMS. The DBMS stores the information in various forms and provides an efficient management and a fast search of information.

Note that using the DBMS one can pose the query and retrieve the required information about the books. An interesting question is " Is it possible to retrieve the relevant and useful information i.e the books referred by most of the users who had referred the book with respect to which the query is posed ?". The Answer is "Data Mining"

Data Mining (DM) is considered to be an important step in the process of knowledge discovery that emphasizes the cleaning, warehousing, mining of knowledge in data bases. It is a form of artificial intelligence that uses automated processes to find information. Although its use in libraries is limited, data mining has been used successfully for several years in the scientific, medical and business communities for tracking behavior of individuals and groups, processing medical information and a number of other aplications.

1. Definitions

1. the non trivial extraction of implicit, primarily unknown and potentially useful information from data.
2. 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 always voluminous, but as it stands of low value as no direct use can be made to it; it is the hidden information in the data that is useful.
3. Webopedia defines the D M as " A class of data base applications that look for hidden patterns in a group of data that can be used to predict future behavior. For example data mining software can help retail companies. The term is commonly misused to describe software that presents data in new ways. True data mining software does not just change the presentation, but actually discovers previously unknown relationships among the data"

2. Data Mining : The New Paradigm

Over the past three decades computers have been used to capture details of business transactions such as banking and credit card records, retail sales, manufacturing warranty and telecommunications etc.

To distill information form a database we obviously need to perform analysis at some time. The key question is *when* In other words; does the analysis take place at the time user needs the knowledge or is it done before hand ,with knowledge ready to access?.

There are two distinct paradigms for empowering users with knowledge.

a. Data Analysis Paradigm:- In this users operate on data to discover information. This paradigm relies on the analysis on demand approach.
b. Knowledge Access Paradigm:- In this analysis is automatically done before hand, refined patterns are pre-generated and users just get knowledge when needed. It provides a multitude of benefits to the user.

- ✍ Condensed information.
- ✍ Easy to use, yet powerful.
- ✍ Fast response and overall efficiency.
- ✍ Accuracy and quality.
- ✍ Up-to-date knowledge.

Comparison:- Data mining and DBMS.

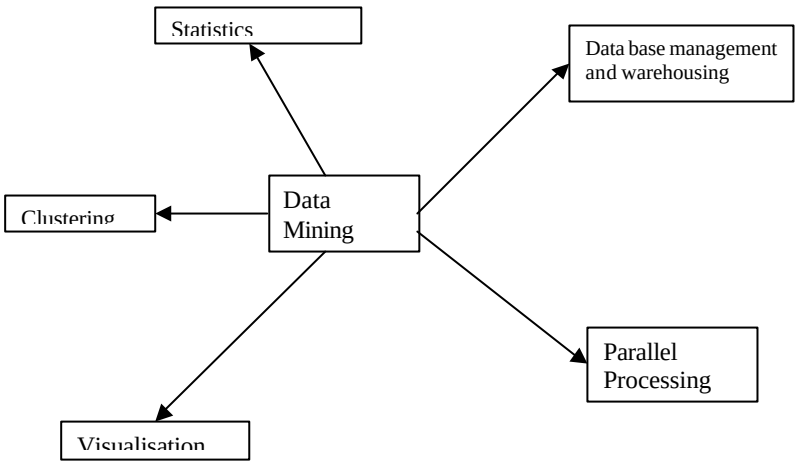
- ✍ DBMS- Queries based on the data held eg.
  Last month's sales for each product.
  Sales grouped by customer age etc.
  List of customers who lapsed their policy.

- ✍ Data Mining- Infer knowledge from the data held to answer queries eg.
  what characteristics do customers share who lapsed their policies and how do they differ from those renewed their policies.

3. Data Mining Techniques

Data Mining encompassed a number of technical approaches such as:-

- ✍ Data base management and ware housing.
- ✍ Statistics .
- ✍ Clustering.
- ✍ Visualization.
- ✍ Parallel processing etc.



Data Mining is the analysis of data and the use of software techniques for finding patterns and regularities in sets of data. It can be distinguished from other technologies in that it makes choices and calculations for the searcher and then categorizes information based on those choices.

4. Data Mining Applications

Data Mining research is being carried out in various disciplines .

- Medicine : drug side effects, hospital cost-analysis, genetic sequence analysis, prediction etc.
- Finance : Stock market prediction, credit assessment, fraud detection etc.
- Marketing \sales: Product analysis buying patterns, sales prediction target mailing, identifying unusual behavior etc.
- Knowledge Acquisition.
- Scientific discovery- Superconductivity research etc.
- Engineering- Automotive diagnostic expert systems, fault detention etc.

5. Data Mining process.

1. Data preprocessing:

   - Heterogeneity resolution.
   - Data cleansing.
   - Data warehousing.

2. Data Mining Tools Applied

- User bias i.e can direct Data Mining tools to areas of interest.
- Attributes of interest in databases.
- Goal of discovery.
- Domain knowledge.
- Prior knowledge or belief about the domain.

6. Data Mining Problems / Issues

- Noisy data.
- Missing values/Incomplete data.
- Static data/ sparse data.
- Dynamic data.
- Relevance.
- Heterogeneity.
- Size and complexity of data.

7. Application of Data Mining is LIS

Data Mining is an inter disciplinary field driven by applications. It involves techniques for machine learning, pattern recognition, statistics, linguistics and visualization with the rapid expansion of full text and bibliographic database on the web. Internet navigation has become a serious concern to libraries. The application of data mining tools in LICs is in its infancy. The data mining techniques can be considered for application in the following areas is LICs.

- Data mining is more suitable to libraries that purchase access to full text databases rather than physical materials and bibliographic databases.
- Users can apply the techniques to measure the use patterns and reuse patterns of databases and software.
- In the area of bibliometrics to discover patterns in unidentified knowledge areas.
- Data Mining techniques can be applied to text analysis tasks such as discipline extractions.
- Data Mining techniques can be in corporated for Information Retrieval process for better browsing and searching

8. References

1. Adrianns P and Zantiuge P (1996) Data Mining. Mass: Addison-Wesly.
2. Rajagopalan B and Krovi R (2002) Benchmarking data mining algorithms. Journal of Database Management 13(1) 25-35.
3. http://management webopaedia.com/ term/D/

About Authors

Mahesh Mudhol is a senior Lecturer in Library & Information Science in University of Mangalore. He has vast experience of teaching. He has published several number of papers in national conferences & seminars.

Purushothama Gowda is working as Assistant Librarian in University of Mangalore. He has published number of papers in seminar & conferences