International Journal of Advanced Computer Research (ISSN (print): 2249-7277 ISSN (online): 2277-7970)
Volume-2 Number-4 Issue-6 December-2012

Data Mining, Applications and Knowledge Discovery

Neha Purohit[1], Sapna Purohit[2], Ritesh Kumar Purohit[3]
[1]Department of Computer Science & Engineering Medicaps Institute of Technology and Management Indore, India
[2]Department of Media & System Engineering Military college of Telecommunication and Engineering Mhow, Indore, India
[3]Department of Computer Science & Engineering Shri Govindram Seksaria Institute of Technology and Science Indore, India

Abstract

This paper explores about the mining of data and finding essential information from huge amounts of data. Extracting or "mining" knowledge from large amounts of data is known as Data Mining. Use of algorithms to extract the information and patterns is derived by the KDD process. Knowledge Discovery in Databases is a process of finding useful information and patterns in data. Research in data mining continues growing in business and in learning organization over coming decades.

Keywords

Data Mining1, Knowledge Discovery in Databases (KDD) 2, Patterns3, Knowledge Management4

1. Introduction

Across a wide variety of fields, data are being collected and accumulated at a dramatic pace. There is an urgent need for a new generation of computational theories and tools to assist humans in extracting useful information (knowledge) from the rapidly growing volumes of digital data. These theories and tools are the subject of the emerging field of knowledge discovery in databases (KDD).

Data mining is a step in the KDD process that consists of applying data analysis and discovery algorithms that produce a particular enumeration of patterns (or models) over the data. The basis of data mining is a process of using tools to extract useful knowledge from large datasets; data mining is an essential part of knowledge management and point that data mining can be useful for KM in two main manners: (i) to share common knowledge of business intelligence (BI) context among data miners and (ii) to use data mining as a tool to extend human knowledge. Thus, data mining tools could help organizations to discover the hidden knowledge in the enormous amount of data.

2. Why Do We Need KDD?

The traditional method of turning data into knowledge relies on manual analysis and interpretation. For example, in the health-care industry, it is common for specialists to periodically analyze current trends and changes in health-care data, say, on a quarterly basis. The specialists then provide a report detailing the analysis to the sponsoring health-care organization; this report becomes the basis for future decision making and planning for health-care management.

There is an urgent need for a new generation of computational theories and tools to assist humans in extracting useful information (knowledge) from the rapidly growing volumes of digital data. As databases grow larger, decision-making from the data is not possible; need knowledge derived from the stored data.

For this (and many other) applications, this form of manual probing of a data set is slow, expensive, and highly subjective. In fact, as data volumes grow dramatically, this type of manual data analysis is becoming completely impractical in many domains.
Databases are increasing in size in two ways: (1) the number N of records or objects in the database and (2) The number d of fields or attributes to an object. Databases containing on the order of $N = 109$ objects are becoming increasingly common, for example, in the astronomical sciences. Similarly, the number of fields d can easily be on the order of 102 or even 103, for example, in medical diagnostic applications. Who could be expected to digest millions of records, each having tens or hundreds of fields? We believe that this job is certainly not one for humans.

The need to scale up human analysis capabilities to handling the large number of bytes can be done by KDD, as it is an attempt to address a problem that the digital information era made a fact of life for all of us: data overload.

3. Understanding KDD

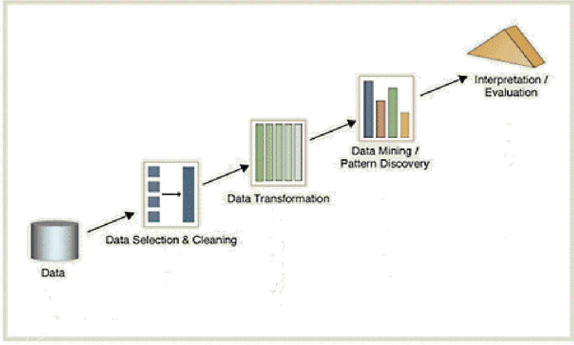
Interpretation / Evaluation
Data Mining / Pattern Discovery
Data Transformation
Data Selection & Cleaning
Data

Figure 1: KDD Process

KDD process consists of iterative sequence methods as follows:

1. Data: Selecting data relevant to the analysis task from the database.
2. Data Selection and Cleaning (Preprocessing): Removing noise and inconsistent data; combining multiple data sources.
3. Data Transformation: Transforming data into appropriate forms to perform data mining.
4. Data mining: Choosing a data mining algorithm which is appropriate to pattern in the data; extracting data patterns.
5. Interpretation/Evaluation: Interpreting the patterns into knowledge by removing redundant or irrelevant patterns; translating the useful patterns into terms that human understandable.

Literature Review:

Data Mining Tasks

Fayyad et.al. (1996) define five main functions of data mining:

1. Classification is finding models that analyze and classify a data item into several predefined classes.
2. Regression is mapping a data item to a real-valued prediction variable.
3. Clustering is identifying a finite set of categories or clusters to describe the data.
4. Dependency Modeling (Association Rule Learning) is finding a model which describes significant dependencies between variables.
5. Deviation Detection (Anomaly Detection) is discovering the most significant changes in the data
6. Summarization is finding a compact description for a subset of data.

Advantages:

1. The analysis and dependency of any variable can be done.
2. The clustering of finite sets if data is the biggest advantage of data mining.

Disadvantages:

1**. Privacy Issues**: Personal privacy has always been a major concern in this country. In recent years, with the widespread use of Internet, the concerns about privacy have increase tremendously. Because of the privacy issues, some people do not shop on Internet. They are afraid that somebody may have access to their personal information and then use that information in an unethical way; thus causing them harm.

2**. Security Issues**: Although companies have a lot of personal information about us available online, they do not have sufficient security systems in place to protect that information. For example, recently the Ford Motor credit company had to inform 13,000 of the consumers that their personal information including Social Security number, address, account number and payment history were accessed by hackers who broke into a database belonging to the Experian credit reporting agency.

3. **Not Accurate**: Trends obtain through data mining intended to be used for marketing purpose or for some other ethical purposes, may be misused. Unethical businesses or people may used the information obtained through data mining to take advantage of vulnerable people or discriminated against a certain group of people. In addition, data mining technique is not a 100 percent accurate; thus mistakes do happen which can have serious consequence.

3. Data Mining

The actual data mining task is the automatic or semi-automatic analysis of large quantities of data to extract previously unknown interesting patterns.

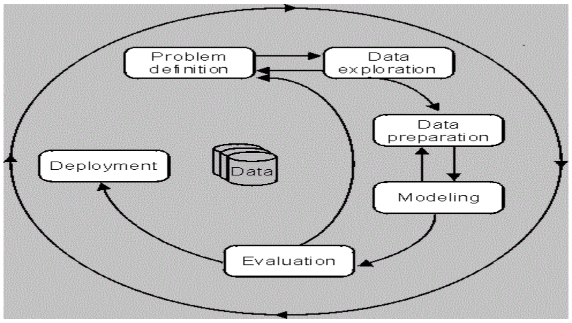

Figure 2: The DM Process Model

Data mining is an iterative process that typically involves the following phases:

Problem definition
A data mining project starts with the understanding of the business problem. Data mining experts, business experts, and domain experts work closely together to define the project objectives and the requirements from a business perspective. The project objective is then translated into a data mining problem definition.
In the problem definition phase, data mining tools are not yet required.

Data exploration
Domain experts understand the meaning of the metadata. They collect, describe, and explore the data. They also identify quality problems of the data. A frequent exchange with the data mining experts and the business experts from the problem definition phase is vital.
In the data exploration phase, traditional data analysis tools, for example, statistics, are used to explore the data.

Data preparation
Domain experts build the data model for the modeling process. They collect, cleanse, and format the data because some of the mining functions accept data only in a certain format. They also create new derived attributes, for example, an average value.
In the data preparation phase, data is tweaked multiple times in no prescribed order. Preparing the data for the modeling tool by selecting tables, records, and attributes, are typical tasks in this phase. The meaning of the data is not changed.

Modeling
Data mining experts select and apply various mining functions because you can use different mining functions for the same type of data mining problem. Some of the mining functions require specific data types. The data mining experts must assess each model.
In the modeling phase, a frequent exchange with the domain experts from the data preparation phase is required.
The modeling phase and the evaluation phase are coupled. They can be repeated several times to change parameters until optimal values are achieved. When the final modeling phase is completed, a model of high quality has been built.

Evaluation
Data mining experts evaluate the model. If the model does not satisfy their expectations, they go back to the modeling phase and rebuild the model by changing its parameters until optimal values are achieved. When they are finally satisfied with the model, they can extract business explanations and evaluate the following questions:

1. Does the model achieve the business objective?
2. Have all business issues been considered?

At the end of the evaluation phase, the data mining experts decide how to use the data mining results.

Deployment
Data mining experts use the mining results by exporting the results into database tables or into other applications, for example, spreadsheets.

4. Data Mining and Knowledge Discovery in the Real World

A large degree of the current interest in KDD is the result of the media interest surrounding successful KDD applications, for example, the focus articles within the last two years in Gadgets, Politics and other large-circulation periodicals. Unfortunately, it is not always easy to separate fact from media hype. In business, main KDD application areas includes marketing, finance (especially investment), fraud detection, manufacturing, telecommunications, and Internet agents.

Marketing: In marketing, the primary application is database marketing systems, which analyze customer databases to identify different customer groups and forecast their behavior. Business Week (Berry 1994) estimated that over half of all retailers are using or planning to use database marketing, and those who do use it have good results; for example, American

Express reports a 10 to 15 percent increase in credit-card use. Another notable marketing application is market-basket analysis (Agrawal et al. 1996) systems, which find patterns such as, “If customer bought X, he/she is also likely to buy Y and Z.” Such patterns are valuable to retailers.

Investment: Numerous companies use data mining for investment, but most do not describe their systems. One exception is LBS Capital Management. Its system uses expert systems, neural nets, and genetic algorithms to manage portfolios totaling $600 million; since its start in 1993, the system has outperformed the broad stock market (Hall, Mani, and Barr 1996).

Fraud detection: HNC Falcon and Nestor PRISM systems are used for monitoring credit card fraud, watching over millions of accounts. The FAIS system (Senator et al. 1995), from the U.S. Treasury Financial Crimes Enforcement Network, is used to identify financial transactions that might indicate money laundering activity.

Manufacturing: The CASSIOPEE troubleshooting system, developed as part of a joint venture between General Electric and SNECMA, was applied by three major European airlines to diagnose and predict problems for the Boeing 737. To derive families of faults, clustering methods are used. CASSIOPEE received the European first prize for innovation.

Business intelligence: Data mining can provide a number of significant benefits to an organization. Through business intelligence data mining companies can create descriptive profiles of key business metrics. For example, through business intelligence data mining a company might build the profile of a customer who is likely to pay beyond 120 days to help identify late paying customers before they become later payers. Through business intelligence data mining companies can identify high risk to the organization and put in place strategies and policies that can help minimize or often even eliminate that risks.

While the use of business intelligence data mining in small businesses has not been prevalent over the years due to cost concerns, there are new tools in the market like EMANIO's Insight! business intelligence data mining product that can provide most of the benefits of higher end tools at substantially reduced prices.

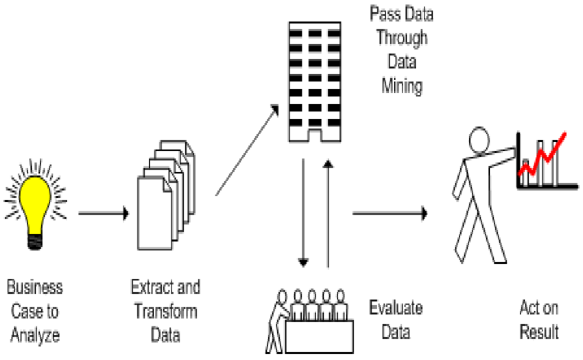

Figure 3: Business Intelligence

Educational Data Mining (called EDM): It is an emerging discipline, concerned with developing methods for exploring the unique types of data that come from educational settings, and using those methods to better understand students, and the settings which they learn in. A key area of EDM is mining computer logs of student performance. Another key area is mining enrollment data. Key uses of EDM include predicting student performance, and studying learning in order to recommend improvements to current educational practice. EDM can be considered one of the learning sciences, as well as an area of data mining. A related field is learning analytics.

5. Problem Domain

Many issues still need to be addressed to reap quality knowledge from the sophisticated algorithms available for data mining. For example:

- How good is the quality of discovered knowledge?
- Does the same method always produce the same results?

Data is an important issue. Dealing with incomplete raw data or erroneous input is not a trivial task. The size of the data set needed to apply an algorithm, duplicate data, and temporal data as well as multimedia representation of data are concerns. How a data mining technique can learn to improve itself through experience is another interesting issue to be considered.

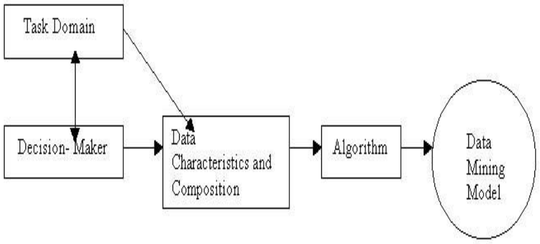

Figure 4: Data Mining Research Framework

Marketing and business intelligence can be the research issues which relates to data integrity. A key implementation challenge is integrating conflicting or redundant data from different sources. For example, a bank may maintain credit cards accounts on several different databases. The addresses (or even the names) of a single cardholder may be different in each. Software must translate data from one system to another and select the address most recently entered.

Another issue is Mining Complex Knowledge from Complex Data.

Data that are not i.i.d. (independent and identically distributed) and many objects which are not independent of each other, and are not of a single type E.g.: interlinked Web pages.

Acknowledgment

This paper has benefited from conversations with many different people far more than can be acknowledged completely here. We thank Prof. O.P Vyas, IIIT-A and Lect. Rahul Rathore for their valuable suggestions and ideas.

References

[1] Tipawan Silwattananusarn and Assoc.Prof. Dr. Kulthida Tuamsuk, "Data Mining and Its Applications for Knowledge Management: A Literature Review from 2007 to 2012", International Journal of Data Mining & Knowledge Management Process (IJDKP), Vol.2, No.5, Thailand, September 2012.

[2] Pardeep Kumar , Nitin and Vivek Kumar Sehgal and Durg Singh Chauhan, "A benchmark to select data mining based classification algorithms for business intelligence and decision support systems", International Journal of Data Mining & Knowledge Management Process (IJDKP), Himachal Pradesh, Vol.2, No.5, September 2012.

[3] Usama Fayyad, Gregory Piatetsky-Shapiro, Padhraic Symth, "Knowledge Discovery and Data Mining: Towards a Unifying Framework", Second International Conference of Knowledge Management & Data Mining (KDD-96) Process, Vol.2, No. 6, Portland, August 1996.

[4] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth,"From Data Mining to Knowledge Discovery in Databases", American Association for Artificial Intelligence, America, Vol. 4, No.37, FALL 1996.

[5] Agrawal, R., and Psaila, G., "Active Data Mining. In Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD-95)", American Association for Artificial Intelligence, Menlo Park, Vol.1, No. 6, Nov 1996.

[6] Umesh Kumar Pandey, S. Pal, "A Data Mining view on Class Room Teaching Language", IJCSI International Journal of Computer Science Issues, ISSN (Online): 1694-0814, Vol. 8, Issue 2, March 2011.

[7] Heikki Mannila, "Theoretical Frameworks for Data Mining", SIGKDD Explorations, Nokia Research Center, Finland, Vol 1, Issue 2, page 30, January 2000.

[8] SANJAY MADRIA, SOUROV S BHOWMICK, W. -K NG, E. P. LIM, "Research Issues in Web Data Mining", Center for Advanced Information Systems, School of Applied Science, Nanyang Technological University, Singapore, 2008.

Ms. Neha Purohit-Indore (M.P) Pursuing Master of Engineering in (Computer Science & Engineering) from Medicaps Institute of Technology & Management, Rau.

Ms. Sapna Purohit-Indore (M.P) working as a Research Associate on Network Simulation Testbed Project Military College of Telecommunication & Engineering, Mhow.

Mr. Ritesh Kumar Purohit-Indore (M.P) Pursuing Bachelor of Engineering in (Computer Science & Engineering) from Shri Govindram Seksaria institute of Science & Technology.