RESEARCH ARTICLE OPEN ACCESS

Importance of Data Mining with Different Types of Data Applications and Challenging Areas

Ms. Aruna J. Chamatkar[*], Dr. P.K. Butey[**]
* (Department of Electronics & Computer Science, RTM Nagpur University, Nagpur)
** (Department of Computer Science, RTM Nagpur University ,Nagpur-32)

ABSTRACT

Data mining is an increasingly popular set of tools for dealing with large amounts of data, often collected in haphazard fashion with many missing values. This type of huge amount of data's are available in the form of tera- to peta-bytes which has drastically changed in the areas of science and engineering. To analyze, manage and make a decision of such type of huge amount of data there are need to techniques called the data mining which will transforming in many fields. In Data Mining data sets will be explored to yield hidden and unknown predictions which can be used in future for the efficient decision making. Data Mining that involves pattern recognition, mathematical and statistical techniques to search data Warehouses and help the analyst in recognizing significant trends, facts relationships and anomalies. In this paper we discuss the importance of data mining , different challenging areas and application areas in data mining .

Keyword - data integration ,data mining , KDD, knowledge, OLAP

I. INTRODUCTION

Data mining is the extraction of useful patterns and relationships from data sources, such as databases, texts, the web. It has nothing to do however with SQL, OLAP, data warehousing or any of that kind of thing. It uses statistical and pattern matching techniques. The concern in data mining are noisy data, missing values, static data, sparse data, dynamic data, relevance, interestingness, heterogeneity, algorithm efficiency, size and complexity of data. The data we have is often vast, and noisy, meaning that it's imprecise and the data structure is complex. This is where a purely statistical technique would not succeed, so data mining is a solution. Data mining has become a popular tool for analyzing large datasets. The efficient database management systems have been very important assets for management of a large corpus of data and especially for effective and efficient retrieval of particular information from a large collection whenever needed. The proliferation of database management systems has also contributed to recent massive gathering of all sorts of information. Information retrieval is simply not enough anymore for decision-making.

II. What are Data Mining and Knowledge Discovery?

With the enormous amount of data stored in files, databases, and other repositories, it is increasingly important, if not necessary, to develop powerful means for analysis and perhaps interpretation of such data and for the extraction of interesting knowledge that could help in decision-making[1].

Data Mining, also popularly known as Knowledge Discovery in Databases (KDD), refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. While data mining and knowledge discovery in databases (or KDD) are frequently treated as synonyms, data mining is actually part of the knowledge discovery process. The following figure (Figure 1.1) shows data mining as a step in an iterative knowledge discovery process.

The Knowledge Discovery in Databases process comprises of a few steps leading from raw data collections to some form of new knowledge. The iterative process consists of the following steps:

1. Data cleaning: It is also known as data cleansing, it is a phase in which noise data and irrelevant data are removed from the collection.
2. Data integration: In this stage, multiple data sources, often heterogeneous, may be combined in a common source.
3. Data selection: At this step, the data relevant to the analysis is decided on and retrieved from the data collection.
4. Data transformation: also known as data consolidation, it is a phase in which the selected data is transformed into forms appropriate for the mining procedure.

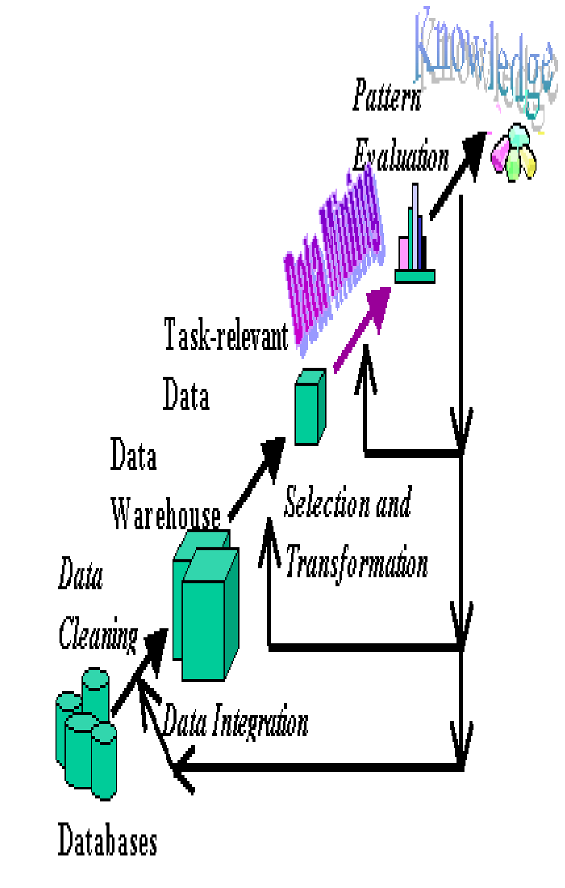

Figure 1.1: Data Mining is the core of Knowledge Discovery process

5. Data mining: it is the crucial step in which clever techniques are applied to extract patterns potentially useful.
6. Pattern evaluation: in this step, strictly interesting patterns representing knowledge are identified based on given measures.
7. Knowledge representation: is the final phase in which the discovered knowledge is visually represented to the user. This essential step uses visualization techniques to help users understand and interpret the data mining results.

It is common to combine some of these steps together. For instance, data cleaning and data integration can be performed together as a pre-processing phase to generate a data warehouse. Data selection and data transformation can also be combined where the consolidation of the data is the result of the selection, or, as for the case of data warehouses, the selection is done on transformed data.

The KDD is an iterative process. Once the discovered knowledge is presented to the user, the evaluation measures can be enhanced, the mining can be further refined, new data can be selected or further transformed, or new data sources can be integrated, in order to get different, more appropriate results[2].

Data mining became the accepted customary term, and very rapidly a trend that even overshadowed more general terms such as knowledge discovery in databases (KDD) that describe a more complete process. Other similar terms referring to data mining are: data dredging, knowledge extraction and pattern discovery.

III. Five Major Elements in Data Mining

1) Extract, transform, and load transaction data onto the data warehouse system.
2) Store and manage the data in a multidimensional database system.
3) Provide data access to business analysts and information technology professionals.
4) Analyze the data by application software.
5) Present the data in a useful format, such as a graph or table.

IV. What can be discovered?

The kinds of patterns that can be discovered depend upon the data mining tasks given. Two types of data mining tasks are there descriptive data mining tasks that describe the general properties of the existing data, and predictive data mining tasks that attempt to do predictions based on inference on available data.

The data mining functionalities and the variety of knowledge they discover are briefly presented in the following list:

- Characterization: Data characterization is a summarization of general features of objects in a target class, and produces what is called characteristic rules. module to extract the essence of the data at different levels of abstractions.
- Discrimination: Data discrimination produces what are called discriminate rules and is basically the comparison of the general features of objects between two classes referred to as the target class and the contrasting class[4,5].
- Association analysis: Association analysis is the discovery of what are commonly called association rules. It studies the frequency of items occurring together in transactional databases, and based on a threshold called support, identifies the frequent item sets.
- Classification: Classification analysis is the organization of data in given classes. Also known as supervised classification, the classification uses given class labels to order the objects in the data collection. Classification approaches normally use a training set where all objects are already associated with known class labels. The classification algorithm learns from the training set and builds a model. The model is used to classify new objects.

- Prediction: Prediction has attracted considerable attention given the potential implications of successful forecasting in a business context. There are two major types of predictions: one can either try to predict some unavailable data values or pending trends, or predict a class label for some data. The latter is tied to classification.
- Clustering: Similar to classification, clustering is the organization of data in classes. However, unlike classification, in clustering, class labels are unknown and it is up to the clustering algorithm to discover acceptable classes.
- Outlier analysis: Outliers are data elements that cannot be grouped in a given class or cluster. Also known as exceptions or surprises, they are often very important to identify.
- Evolution and deviation analysis: Evolution and deviation analysis pertain to the study of time related data that changes in time. Evolution analysis models evolutionary trends in data, which consent to characterizing, comparing, classifying or clustering of time related data.

It is common that users do not have a clear idea of the kind of patterns they can discover or need to discover from the data at hand. It is therefore important to have a versatile and inclusive data mining system that allows the discovery of different kinds of knowledge and at different levels of abstraction. This also makes interactivity an important attribute of a data mining system.

V. What are the kinds of data can be mined ?

- Flat files: Flat files are actually the most common data source for data mining algorithms, especially at the research level. Flat files are simple data files in text or binary format with a structure known by the data mining algorithm to be applied. The data in these files can be transactions, time-series data, scientific measurements, etc.
- Relational Databases: a relational database consists of a set of tables containing either values of entity attributes, or values of attributes from entity relationships. Tables have columns and rows, where columns represent attributes and rows represent tuples. A tuple in a relational table corresponds to either an object or a relationship between objects and is identified by a set of attribute values representing a unique key
- Transaction Databases: A transaction database is a set of records representing transactions, each with a time stamp, an identifier and a set of items. Associated with the transaction files could also be descriptive data for the items.
- Multimedia Databases: Multimedia databases include video, images, audio and text media. They can be stored on extended object-relational or object-oriented databases, or simply on a file system. Multimedia is characterized by its high dimensionality, which makes data mining even more challenging.
- Spatial Databases: Spatial databases are databases that, in addition to usual data, store geographical information like maps, and global or regional positioning. Such spatial databases present new challenges to data mining algorithms.
- Time-Series Databases: Time-series databases contain time related data such stock market data or logged activities. These databases usually have a continuous flow of new data coming in, which sometimes causes the need for a challenging real time analysis. Data mining in such databases commonly includes the study of trends and correlations between evolutions of different variables, as well as the prediction of trends and movements of the variables in time.
- World Wide Web: The World Wide Web is the most heterogeneous and dynamic repository available. Data in the World Wide Web is organized in inter-connected documents[3].

VI. Challenging Problems In Data Mining

1. Developing a Unifying Theory of Data Mining
2. Scaling Up for High Dimensional Data and High Speed Data Streams
3. Mining Sequence Data and Time Series Data
4. Mining Complex Knowledge from Complex Data
5. Data Mining in a Network Setting
6. Distributed Data Mining and Mining Multi-agent Data
7. Data Mining for Biological and Environmental Problems
8. Data-Mining-Process Related Problems
9. Security, Privacy and Data Integrity
10. Dealing with Non-static, Unbalanced and Cost-sensitive Data

VII. Application Of Data Mining

Now a days data mining are used in lots of areas but In this section , here we mainly listed some application areas for data mining[6,7].

1. Data mining Application in Healthcare
2. Future Direction of Health care system through Data mining tools
3. Data mining used in many different areas in manufacturing Engineering
4. Data mining is used for market basket analysis
5. Data mining is used an emerging trends in the educational system
6. Data mining Application can be generic and domain specific
7. Data mining techniques used in the CRM

8. Large scope for application of data mining in Medical Science
9. Data mining Methods are used in the Web Application
10. Data mining method is used to classify the network traffic control
11. Data mining and its techniques is used for an application of Sports Center
12. Data mining methods are used for application in a malicious executable is Threat i.e. in System Security.

VIII. CONCLUSIONS

In this paper , briefly discuss the basic concept related to the data mining, challenging and application areas for data mining. Data mining is more than running some complex queries on the data you stored in your database.. Identifying the format of the information that you need is based upon the technique and the analysis that you want to do.To Analyze, manage and make a decision of such type of huge amount of data we need techniques called the data mining which will transforming in many fields. KDD is the actually the process of finding hidden pattern of the repositories. The different method of data mining are used to extract the pattern and thus the knowledge from this variety databases. Data mining should be applicable to any kind of information repository. The challenges listed by different types of data very significantly. Data Mining methods ,tools and techniques are useful in different application areas.

REFERENCES:

Journal Papers:

[1] M. Shiga, I. Takigawa, and H. Mamitsuka, “A spectral clustering approach to optimally combining numericalvectors with a modular network,” in KDD, *2007, pp. 647–656.*

[2] Osmer R. Zalane,CMPUT 690 principles of knowledge discovery in databases” Introduction to Data mining”.

[3] M.S. Chen.J.Han and P.S. Yu. Data mining : An overview from a database perspective. *IEEE transactions on Knowledge and data engineering 8:866.*

[4] Feldman, Ronen, Will Klosgen, and Amir Zilberstein. “Visualization techniques to explore data mining results for document collections.”, *In Proceedings ofthe Third Annual Conference on KnowledgeDiscovery and Data Mining (KDD), Newport Beach, 1997*

[5] Koperski, J. Adhikary and J. Han, "Spatial Data Mining: Progress and Challenges", SIGMOD'96Workshop on Research Issues in Data Mining and Knowledge Discovery DMKD'96, Montreal,Canada.

[6] N. Padhy, P.Mishra (IJCSEIT) “ The survey of Data mining Application” vol.2 no. 3 June 2012 .

Books:

[7] Han, J. and M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann, 2001.

[8] Introduction to Data Mining with Case Studies by **Gupta G. k**