A Conceptual Overview of Data Mining

B.N. Lakshmi. [#1],G.H. Raghunandhan. [#2]
[#1, 2]*Department of Computer science Engineering, Reva Institute of Technology and Management;*
Bangalore, Karnataka, India
[1]keerthisri.20@gmail.com

***Abstract*—Data mining an non-trivial extraction of novel, implicit, and actionable knowledge from large data sets is an evolving technology which is a direct result of the increasing use of computer databases in order to store and retrieve information effectively .It is also known as Knowledge Discovery in Databases (KDD) and enables data exploration, data analysis, and data visualization of huge databases at a high level of abstraction, without a specific hypothesis in mind. The working of data mining is understood by using a method called modeling with it to make predictions. Data mining techniques are results of long process of research and product development and include artificial neural networks, decision trees and genetic algorithms. This paper surveys the data mining technology, its definition, motivation, its process and architecture, kind of data mined, functionalities and classification of data mining, major issues, applications and directions for further research of data mining technology.**

***Keywords*—data mining, KDD, recent techniques, modeling, further research**.

I.INTRODUCTION

With the advent of computers and means for mass digital storage we started collecting and storing all sorts of data, counting on the power of computers to help sort through this amalgam of information. This massive collections of data stored on disparate structures very rapidly became overwhelming and led to the creation of structured databases and database management systems (DBMS).The database management systems efficiently manage large corpus of data and effective and efficient retrieval of particular information from a large collection whenever needed and also contributes to recent massive gathering of all sorts of information. This retrieval of data as and when needed contributes the technology of data mining. Data mining can be viewed as a result of the natural evolution of information technology. This technology provides a wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge. Data mining is the extraction of interesting patterns or knowledge from huge amount of data. It can be known by different names like knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence and others. The term “data mining” is nothing but analysis of data in a database using tools which look for trends or anomalies without the knowledge of meaning of the data and is primarily used by statisticians, database researchers and business communities. A data mining software does not just change the presentation, but discovers previously unknown relationships among the data. The information on which the data mining process operates is contained in a historical database of previous interactions. In principle, data mining is not specific to one type of media or data. Data mining should be applicable to any kind of information repository. Some kinds of information that is collected are as follows:

A. Business transactions

Every transaction in the business industry is (often) “memorized” for perpetuity. Such transactions are usually time related and can be inter-business or intra-business operations effective use of the data in a reasonable time frame for competitive decision making is definitely the most important problem to solve for businesses that struggle to survive in a highly competitive world.

B.Scientific data

Our society is amassing colossal amounts of scientific data that need to be analyzed. Unfortunately, we can capture and store more new data faster than we can analyze the old data already accumulated.

C.Medical and personal data

From government census to personnel and customer files, very large collections of information are continuously gathered about individuals and groups. This type of data often reveals if the information is collected, used and even shared. When correlated with other data this information can shed light on customer behavior.

D.Surveillance video and pictures

Storing the video tapes and digitizing them for future use and analysis.

E.Satellite sensing

There are a countless number of satellites around the globe and all send a non-stop stream of data to the surface. Many satellite pictures and data are made public as soon as they are received in the hopes that other researchers can analyze them.

F.Text reports and memos (e-mail messages)

Most of the communications are based on reports and memos in textual forms often exchanged by e-mail.

978-1-61284-810-5/11/$26.00 ©2011 IEEE

These messages are regularly stored in digital form for future use and reference creating formidable digital libraries.

G. The World Wide Web repositories

In the World Wide Web documents of all sorts of formats, content and description are collected and inter-connected with hyperlinks making it the largest repository of data. Despite of its dynamic and unstructured nature, its heterogeneous characteristic, and its very often redundancy and inconsistency, the World Wide Web is the most important data collection regularly used for reference because of the broad variety of topics covered and the infinite contributions of resources and publishers.

II.ARCHITECTURE AND PROCESS OF DATA MINING

A. The Architecture of Data Mining

The architecture of a typical data mining system has the following major components:

1) Database, data warehouse, or other information repository: This component is one or a set of databases, data warehouses, spread sheets, or other kinds of information repositories. Data cleaning and data integration techniques may be performed on the data.

2) Database or data warehouse server: The component is responsible for fetching the relevant data, based on the data mining request of the user.

3) Knowledge base: This is the domain knowledge that is used to guide the search, or evaluate the interestingness of resulting patterns. It includes concept hierarchies that are used to organize attributes or attribute values into different levels of abstraction.

4) Data mining engine: This is an essential component of the data mining system and ideally consists of a set of functional modules for tasks such as characterization, association analysis, classification, evolution and deviation analysis.

5) Pattern evaluation module: This component typically employs interestingness measures and interacts with the data mining modules so as to focus the search towards interesting patterns. It can access interestingness thresholds stored in the knowledge base. Alternatively, the pattern evaluation module may be integrated with the mining module, depending on the implementation of the data mining method used. Efficient data mining is possible by pushing the evaluation of pattern interestingness deeply into the mining process so as to connect the search to only the interesting patterns.

6) Graphical user interface: This module communicates between users and the data mining system, and allows the user to interact with the system by specifying a data mining query or task, providing information to help focus the search, and performing exploratory data mining based on the intermediate data mining results. This component also allows the user to browse database and data warehouse schemas or data structures, evaluate mined patterns, and visualize the patterns in different forms.

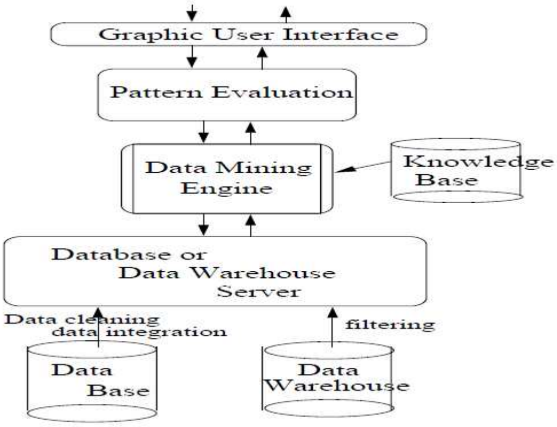

Fig 1: Architecture of a typical data mining system.

B. The Process of Data Mining

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 can also be termed as data cleansing. It is a phase wherein noise data and irrelevant data are removed from the collection.

2) Data integration: In this stage multiple data sources that are heterogeneous can be combined in a common source.

3) Data selection: In this phase the data relevant to the analysis is decided on and retrieved from the data collection.

4) Data transformation: It can also be known as data consolidation and it is a phase in which the selected data is transformed into forms appropriate for the mining procedure.

5) Data mining: This is a crucial phase wherein skillful techniques are applied to extract patterns potentially useful.
Pattern evaluation: In this current phase strictly interesting patterns representing knowledge are identified with respect to the given measures.

6) Knowledge representation: This is the final phase in which the discovered knowledge is visually represented to the user. This is an essential step which uses visualization techniques to help users understand and interpret the data mining results.

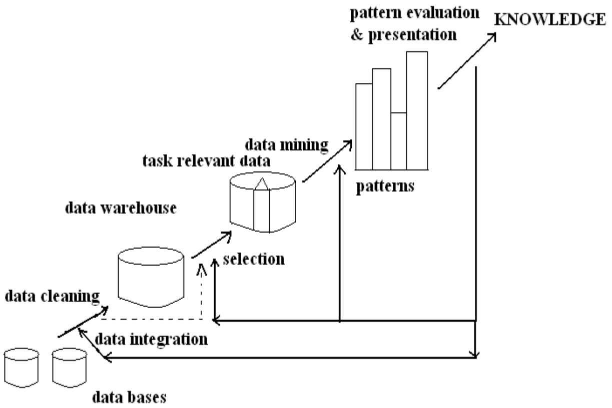

Fig2: Data mining as a process of knowledge discovery

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.

The types of data mined are as follows:

1) Flat files: 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.

2) Relational Databases: Briefly, 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.

3) Data Warehouses: A data warehouse as a storehouse is a repository of data collected from multiple data sources (often heterogeneous) and is intended to be used as a whole under the same unified schema. A data warehouse gives the option to analyze data from different sources under the same roof.

4) 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.

5) 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.

6) Spatial Databases: Spatial databases are databases that store geographical information like maps, and global or regional positioning in addition to usual data.

7) 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. These documents can be text, audio, video, raw data, and even applications. Data mining in the World Wide Web, or web mining, is often divided into web content mining, web structure mining and web usage mining.

III.FUNCTIONALITIES AND CLASSIFICATIONS OF DATA MINING

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

A. Characterization

Data characterization is a summarization of general features of objects in a target class, and produces what is called characteristic rules. The data relevant to a user-specified class are normally retrieved by a database query and run through a summarization module to extract the essence of the data at different levels of abstractions.

B. Discrimination

Data discrimination produces what are called discriminant 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.

C. 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. Another threshold, confidence, which is the conditional probability than an item appears in a transaction when another item appears, is used to pinpoint association rules. Association analysis is commonly used for market basket analysis.

D. 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.

E. Prediction

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 and is tied to classification. Once a classification model is built based on a training set, the class label of an object can be foreseen based on the attribute values of the object and the attribute values of the classes. Prediction is referred to the forecast of missing numerical values, or increase/ decrease trends in time related data. The major idea is to use a large number of past values to consider probable future values.

F. Clustering

It is the organization of data in classes and is similar to classification. In clustering class labels are unknown and it is up to the clustering algorithm to discover acceptable classes. Clustering is also called unsupervised classification as the classification is not dictated by given class labels. There are many clustering approaches based on the principle of maximizing the similarity between objects in same class called intra-class and minimizing the similarity between objects of different classes called inter-class similarity.

G. Outlier analysis

Outliers are data elements that cannot be grouped in a given class or cluster. They are also known as exceptions or surprises and are often very important to identify. Outliers can reveal important knowledge in other domains, can be very significant and their analysis valuable.

H. Evolution and deviation analysis:

This pertains to the study of time related data that change with time. Evolution analysis model's evolutionary trends in data, which consent to characterizing, comparing, classifying or clustering of time related data. Deviation analysis considers differences between measured values and expected values, and attempts to find the cause of the deviations from the anticipated values.

Many data mining systems are available or are being developed, among which some are specialized systems dedicated to a given data source or are confined to limited data mining functionalities and other are more versatile and comprehensive. Data mining systems can be categorized according to various criteria their classification is as follows:

A. Classification according to the type of data source mined

This classification categorizes data mining systems according to the type of data handled such as spatial data, multimedia data, time-series data, text data, World Wide Web.

B. Classification according to the data model drawn on

This classification categorizes data mining systems based on the data model involved such as relational database, object-oriented database, data warehouse, transactional and others.

C. Classification according to the king of knowledge discovered

This classification categorizes data mining systems based on the kind of knowledge discovered or data mining functionalities, such as characterization, discrimination, association, classification, clustering and others. Some systems offer several data mining functionalities together and tend to be comprehensive systems.

D. Classification according to mining techniques used

Different techniques are employed and provided in data mining systems. The classification categorizes data mining systems according to the data analysis approach used such as machine learning, neural networks, genetic algorithms, statistics, visualization, database oriented or data warehouse-oriented and others. The classification can also take into account the degree of user interaction involved in the data mining process such as query-driven systems, interactive exploratory systems, or autonomous systems. A comprehensive system would provide a wide variety of data mining techniques to fit different situations and options, and offer different degrees of user interaction.

IV.BENEFITS AND LIABILITIES OF DATA MINING AND ITS APPLICATIONS

There are many benefits that can be obtained through the application of data mining technology. Some of the benefits of data mining are as follows:

- Helps to unearth facts about customers from your database, which you previously didn't know about, including purchasing behavior.
- Lends automation benefits to existing hardware and software.
- Crediting/Banking: helpful to financial institutions in such areas as loan information and credit reporting.
- Research: makes the process of data analysis faster.
- Law enforcement: can assist law enforcers with keying out criminal suspects and taking them into custody, by looking into trends in various behavior patterns.
- Marketing: helps to foretell the products which customers would like to buy.
- Transportation: to evaluate loading patterns.
- Medicine: to discover effective medical therapies for diverse illnesses.
- Insurance: to make out fraudulent behavior.
- Enhances efficiency and saves money.

Data mining is an emerging trend and ubiquitous and before it develops into a conventional, mature and trusted discipline, many still pending issues have to be addressed. Some of these issues are:

A. Security and social issue

Security is an important issue with any data collection when it is shared and is intended to be used for strategic decision-making. This becomes controversial given the confidential nature of some of this data and the potential illegal access to the information. Data mining could disclose new implicit knowledge about individuals or groups that could be against privacy policies, especially if there is potential dissemination of discovered information. There arises another issue from this concern that is the appropriate use of data mining. Due the competitive advantage attained from implicit knowledge discovered, some of the important information could be withheld and other information can be widely distributed and can be used without control.

B. User interface issues

The knowledge discovered by data mining tools is useful as long as it is interesting, and above all understandable by the user. The major issues related to user interfaces and visualization is "screen real-estate", information rendering, and interaction. Interactivity with the data and data mining results is crucial since it provides means for the user to focus and refine the mining tasks, as well as to picture the discovered knowledge from different angles and at different conceptual levels

C. Mining methodology issues

These issues pertain to the data mining approaches applied and their limitations. More than the size of data, the size of the search space is even more decisive for data mining techniques. The size of the search space is often depending upon the number of dimensions in the domain space. The search space usually grows exponentially when the number of dimensions increases. This is known as the curse of dimensionality. This "curse" affects so badly the performance of some data mining approaches that it is becoming one of the most urgent issues to solve.

D. Performance issues

Many artificial intelligence and statistical methods are there for data analysis and interpretation and are often not designed for the very large data sets data mining deals with. This raises the issues of scalability and efficiency of the data mining methods when processing considerably large data. Other topics in the issue of performance are incremental updating, and parallel programming.

E. Data source issues

There are many issues related to the data sources, some are practical such as the diversity of data types, while others are philosophical like the data glut problem. Heterogeneous data sources, at structural and semantic levels, pose important challenges not only to the database community but also to the data mining community.

Some of the applications of data mining are:

A. Data Mining in Agriculture

Recent technologies are nowadays able to provide a lot of information on agricultural-related activities, which can then be analyzed in order to find important information

B. Surveillance / Mass surveillance

Surveillance is the monitoring of the behavior, activities, or other changing information, usually of people and often in a surreptitious manner. It most usually refers to observation of individuals or groups by government organizations, but disease surveillance, for example, is monitoring the progress of a disease in a community.

C. National Security Agency

The National Security Agency/Central Security Service (NSA/CSS) is a crypto logic intelligence agency of the United States Department of Defense responsible for the collection and analysis of foreign communications and foreign signals intelligence, as well as protecting U.S. government communications and information systems which involves cryptanalysis and cryptography.

1) Quantitative structure-activity relationship: is the process by which chemical structure is quantitatively correlated with a well defined process, such as biological activity or chemical reactivity.

2) Customer analytics: Customer analytics is a process by which data from customer behavior is used to help make key business decisions via market segmentation and predictive analytics.

3) Police-enforced ANPR in the UK: The UK has an extensive automatic number plate recognition (ANPR) CCTV network. Police and security services use it to track UK vehicle movements in real time. The resulting data are stored for 5 years in the National ANPR Data Centre to be analyzed for intelligence and to be used as evidence.

4) Stellar wind (code name): Stellar Wind is the open secret code name for certain information collection activities performed by the United States' National Security Agency.

V. CONCLUSION

Data mining is a technique that offers great promise in helping organizations uncover patterns hidden in their data that can be used to predict the behavior of customers, products and processes. However, data mining tools need to be guided by users who understand the business, the data, and the general nature of the analytical methods involved. Realistic expectations can yield rewarding results across a wide range of applications, from improving revenues to reducing costs. Regarding the practical issues related to data sources, there is the subject of heterogeneous databases and the focus on diverse complex data types. We are storing different types of data in a variety of repositories. It is difficult to expect a data mining system to effectively and efficiently achieve good mining results on all kinds of data and sources. Different kinds of data and sources may require distinct algorithms and methodologies. Currently, there is a focus on the motivation or the need for data mining. We have given a brief explanation about the typical architecture of data mining and explained the steps of the data mining process. This paper abstracts the functionalities of data mining and describes the classification of data mining systems. It spills the lime light on the benefits the data mining technology offers to the present day world. We also discus about the major issues that need to be addressed and mention a few applications wherein data mining technology can be applied. Therefore, from a strategic perspective, the need to navigate the rapidly growing universe of digital data will rely heavily on the ability to effectively manage and mine the raw data.

REFERENCES

[1] Han J. and M. Kamber (2000), Data Mining: Concepts and Techniques, Academic Press, San Diego, CA.

[2] Introduction to Data Mining and Knowledge Discovery, Third Edition by Two Crows Corporation.

[3] DATA MINING: ACONCEPTUAL OVERVIEW
Joyce Jackson, Management Science Department, University of South Carolina, joyce.jackson@sc.edu

[4] M. S. Chen, J. Han, and P. S. Yu. Data mining: An overview from a database perspective. IEEE Trans. Knowledge and Data Engineering, 8:866-883, 1996.

[5] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy.Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, 1996.

[6] Tan, Steinbach, Kumar “Introduction to Data Mining”

[7] Data Mining: Introductory and Advanced Topics Margaret Dunham