A Review: Data Mining, its Issues, Functionalities and Applications

[1]Sandeepak Bhandari, [2] Tarun Sharma, [3] Jagpreet Singh
CT Institute of Technology & Research
Greater Kailash, G.T Road, Maqsudan
Jalandhar, www.ctgroup.in

Abstract

Data mining can be defined as process of extracting knowledge from huge amount of data such as Data warehouses, Databases etc. Data mining help to take efficient and effective decision. Data mining is used in various fields including Retail Industry, Sales/Marketing, Health Care and Insurance, Intrusion Detection etc. But Data Mining is suffered from various issues such as Mining methodology and user interaction issues, Performance issues, Issues related to the diversity of database types as well as challenges for Data Mining

Keywords:-

Data Mining Architecture, Functionalities, Issues, Applications and Challenges in Data Mining.

Introduction: -

Data mining can be defined as the process of extracting or mining information and knowledge from large amount of data. The growth of information technology in various fields lead to growth of large volume of data and storage of data in various formats like, documents, images, sound recordings, videos, scientific data, records and many new data formats. The data is collected from various fields need appropriate mechanism for extracting knowledge and information from large volume of data, so that a better decision can be made. This process refers to Knowledge Discovery in Database (KDD), and often called Data Mining. The information gained from large repositories can be used for various applications market

Analysis, and customer retention, to production control, fraud detection. One of the core functions of Data Mining is to use various methods and algorithms for finding out patterns from stored data.

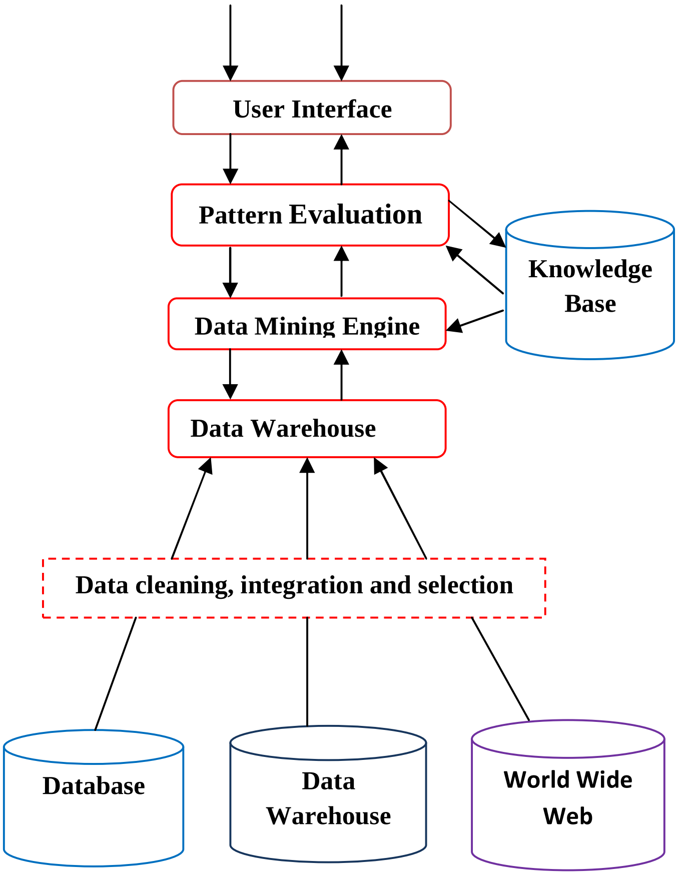
Various factors which lead to growth in the field of data mining as well as knowledge Discovery are as follows:-

The rapid growth in Computing Capability and Storage Capacity.

- Use of Multimedia Data in large volume.
- The availability of increased access to data from Web navigation and intranets i.e. use of internet.
- Storage of Data in Data Warehouse so that enterprise can access data.
- To make a better decision strategy.

Architecture of Data Mining System:-

Architecture of a typical Data Mining System as shown in below figure. Data Mining's architecture consists of many components namely User Interface, Pattern Evaluation, Data Mining Engine, Data Warehouse Server, and Knowledge Base.



Knowledge Base:-It is a domain knowledge that can be used to guide the search or to evaluate the pattern. Such knowledge contain concept of hierarchies, can be used to organize attributes or its values in different levels of abstraction.

Data Mining Engine:-It is an essential component of data mining system and comprises sets of functional components that performs various tasks namely characterization, association and correlation analysis, classification, clustering, prediction and many more.

Pattern Evaluation Module:-This component typically performs interestingness measures and has to interact with the data mining engine module to search interesting pattern.

User Interface:-User Interface module communicates between data mining system and user. It allow the user to communicate and interact with the system by specifying their query, providing information help to focus on search and performing exploratory data mining based on the intermediate data mining results.

Data Mining Functionalities:-

Data mining functionalities can be used to specify what type of patterns can be found in stored data is major task of Data Mining. Data Mining tasks can be divided into two major Categories:-

- Descriptive and
- Predictive

Descriptive: - Descriptive mining tasks characterize the general properties of the data in the database.
Predictive: -Predictive mining tasks perform inference on the current data in order to make predictions.

Data mining functionalities and the kinds of patterns recognized by Data Mining are described as follows:-

Concept/Class Description:

Characterization and Discrimination-Data can be associated with concept or class. For instance, in an Electronic store classes of items for sale includes computers, printers, scanners and concepts of customers include big Spenders and budget Spenders. It can be beneficial to describe individual classes and concepts in concise terms. Such Description of a class is known as class/concept description.
Data characterization means summarizing the data of the class under study, generally known as target class. Data discrimination is comparison between the target class and one or set of comparative class, generally known as contrasting class.

Classification and Prediction:-Classification can be defined as the process of finding a model (or a method) that define and differentiate data concepts or class, so that it can be used for the purpose to predict the class of objects whose class is unknown. The new derived model is based on the data object whose class is known, and is called training data. The new derived model can be represented in various forms, such as Decision trees, neural networks, Classification Rules namely *IF-THEN* and many more. A decision tree is like a tree structure, where each and every node denotes a test on attribute and each branch denote an output of the test and leaves denote classes. As there are various methods for classification model, such as support vector machines, and *k*-nearest neighbour classification. The term prediction refers to both class label prediction and numeric prediction. Regression analysis is strategy that is used for numeric prediction.

Mining Frequent Patterns, Associations, and Correlations:-

Frequent patterns can be defined as the pattern that occurs frequently in data. There are numbers of frequent patterns, including subsequences, substructures and item sets that occur frequently. A frequent item sets can be defined as the set of items that appear together in a transactional data set frequently, for instance Milk and Bread, Shampoo and Conditioner. A frequent subsequence's pattern can be defined as the pattern that customers tend to purchase first. For instance first a PC, followed by a digital camera and then a memory card is a Sequential pattern. A substructure can be defined as different structural forms such as trees, graphs. If a substructure occurs frequently, \then it is known as Structural pattern.

Cluster Analysis:-

Unlike Prediction and Classification which can analyse the class labelled data objects, Whereas Clustering analyse the data objects whose class labels are unknown. As class labels are not present in the training data because they are not known to begin with. The objects are clustered are based on the maximizing the infraclass similarity and

Minimizing the interclass similarit*y*. Its means that clusters of objects are formed in such a way that there is maximum similarity between objects within a cluster and but there is maximum dissimilarity between objects of different clusters.

Outlier Analysis:-Some database contains objects that do not compile with general behaviour and model of data. Such objects are known as Outliers. Generally, most of data mining methods discard outliers as noise or exceptions. The analysis of outliers is known as Outlier Mining. Outliers can be detected with the help of statistical tests that assume probability model for the data

Issues in Data Mining:-

As there are various issues in data mining, but some of major issues in data mining can be broadly divided into three categories namely User Interaction, Performance and Diverse Data types.

Mining methodology and user interaction issues: These refer to mining of knowledge from data at multiple granularities, and use of specific field knowledge and Visualization.

Mining different kinds of knowledge in databases:-As different Users want different types of information, so data miming should cover a wide range of spectrum and knowledge discovery task including prediction, clustering, outlier analysis, correlation analysis.

Interactive mining of knowledge at multiple levels of abstraction: It is difficult to know what exactly to be discovered in database, so that's why data mining process should be interactive. An Interactive Data Mining System allow user to focus in pattern to be search, providing and refining data mining requests based on returned results.

Incorporation of background knowledge: Background Knowledge or information regarding the field or domain under study must be used in knowledge discovery process. It allows discovered process to be express in concise terms and at different levels of abstraction. Knowledge related to database such as integrity constraints and deduction rules can help focus and speed up a data mining process

Handling noisy or incomplete data: Database contains data may reflect noise**,** incomplete data objects or exceptional cases. When mining data regularities, these Objects may cause confuse the process, causing the knowledge model constructed to over fit the data. Which cause discovered process to be poor?

Performance issues:-
The performance issues include efficiency, scalability, and parallelization of data Mining algorithms.

Efficiency and scalability of data mining algorithms:-As there is huge amount of data in database, to extract information from database an effective and scalable algorithm is required. Froma database perspective on knowledge discovery, efficiency and scalability are key issues in the implementation of data mining systems.

Parallel, distributed, and incremental mining algorithms:-
Due to very large size of databases, wide distribution of data and the computational complexity of various data mining methods are the factors which lead to the development of parallel and distributed data mining algorithm. These algorithms split the data into partitions which are process in parallel. The results from different partitions then combined together. Due to high cost of data mining processes which lead to the need of incremental data mining algorithm that incorporate database updates without having to mine the entire data again "from scratch."

Issues related to the diversity of database types:-
Handling of relational and complex types of data:- Relational databases and data warehouses are commonly used coefficient and effective data mining system for such data is important. As other database contains complex data such as hypertext, multimedia data, spatial data, temporal data and transactional data. It is impossible for one system to contain all kinds of data, and gives the diversity of data types and different aim of data mining. Specific data mining systems should be constructed for mining specific kinds of data.

Mining information from heterogeneous databases and global information systems: Computer Networks such as Local Area Network (LAN) and Wide Area Network (WAN) interconnect number of sources of data and information which lead to formation of large, distributed and heterogeneous databases. The formation of information and knowledge from different sources of data like semi structured, structured or unstructured data with diverse semantics contain a great challenge to data mining. Data mining can help disclose high level data regularities in many heterogeneous databases that can be discovered by a simple query system and can improve information exchange and interoperability in heterogeneous Databases.

Applications of Data Mining:-

Data Mining for Financial Data Analysis:-There are many banks and financial institutions which offer large varieties of banking services namely Checking and Saving Accounts, credit and investment services and many offers insurance services and stock investment services. The data (Financial) collected in banking and financial industries are relatively complete, reliable, and of high quality, which provides systematic data analysis and data mining. Some of the cases includes: Design and construction of data warehouses for multidimensional data analysis and data mining, Loan payment prediction and customer credit policy analysis, Classification and clustering of customers for targeted marketing and many more.

Data Mining for the Retail Industry:-

Retail industry is a major application area for data mining. Its mean that data mining collects large amount of data on sales, customer shopping history, goods transportation, consumption and services. The quantity of data collected by data mining from various fields of retail industry expand continuously and at fast rate., due to increasing of ease, availability and popularity of business by advertisement on internet and e-commerce. Today, there are many On-Line Shopping Sites which provides facilities for customers to do shopping by sitting at home with the help of internet. Some of popular On-lone shopping sites are .Myntra.com. Amazon.com and many more. Retail data mining can help to identify customer buying behaviours activities, to identify customer shopping patterns and trends, can improve the quality of customer service, provide customer satisfaction, achieve better customer retention and satisfaction, enhance goods consumption ratios, design more effective goods transportation and distribution policies, and reduce the cost of business. Here are some examples of data mining in retail industry are Multidimensional analysis of sales, customers, products, time, and region, Analysis of the effectiveness of sales campaign.

Data mining for the Telecommunication Industry:-The Telecommunication industry provides services for both Local and Long Distance telephone services to facilitate many other beneficial communication services including fax, pager, cellular phone, Internet messenger, images, e-mail, computer and Web data transmission, and other data traffic. The interaction of computer network, internet, and many ways of communication is underway. With the development of technology and in computer industry, by which , the telecommunication market is rapidly expanding and highly competitive, Which leads to great demand for data mining in telecommunication industry to understand the business involved, identify telecommunication patterns, catch fraudulent activities, make better use of resources, and improve the quality of service.

Data Mining Applications in Sales/Marketing:-

Data Mining helps businesses to know and understand the hidden patterns which are hidden inside the historical purchasing transaction data, which help in planning, organizing, managing and launching new market in a cost effective way. Data Mining can be used for Basket Analysis so that information can be provided on what product combinations were purchased together, when they were bought and in what sequence. This information helps businesses promote their most profitable products and maximize their profit.

Data Mining Applications in Health Care and Insurance:-

The development of insurance industries are entirely depends on the capability of converting data into information about customer, competitors and its market**.** Data mining is used in insurance industry lately but brought tremendous competitive advantages to the companies who have implemented it successfully. The data mining applications in insurance industry are listed below: Data mining is applied in claims analysis such as identifying which medical procedures are claimed together. Data mining enables to forecasts which customers will potentially purchase new policies. Data mining allows insurance companies to detect risky customers' behaviour patterns. Data mining helps detect fraudulent behaviour.

Data Mining for Intrusion Detection:-

Due to rapid growth of internet and increasing availability of tools and tricks for intruding and attacking networks have prompted intrusion detection to become a critical component of network administration. An intrusion can be defined as any set of actions that can threaten the

Integrity, confidentiality, or availability of a network resource (such as user accounts, file systems, system kernels, and so on).There are the areas in which data mining technology can be applied for intrusion detection which includes Development of data mining algorithms for intrusion detection, Association and correlation analysis, and aggregation to help select and build discriminating attributes, Analysis of stream data, Distributed data mining, Visualization and querying tools.

Challenges in Data Mining:-

Distributed data: The data to be mined or extract is stored in distributed computing environments on different platforms. Both for technical and for organizational reasons it is not feasible to bring all the data to a centralized place. So, development of algorithms, tools, and services is required that provides the facilities for mining of distributed data.

Distributed operations: More and more data mining operations and algorithms will be available on the grid in future. To provide facilities for seamless integration of these resources into distributed data mining systems for complex problem solving, novel algorithms, tools, grid services and other IT infrastructure need to be developed.

Massive data: For mining large, massive and high-dimensional data sets (out-of-memory, parallel, and distributed algorithms) development of algorithm is needed. Complex data types: Increasingly complex data sources, structures, and types (like natural language text, images, time series, multi-relational and object data types etc.) are emerging. Grid-enabled mining of such data will require the development of new methodologies, algorithms, tools, and grid services.

Data privacy, security, and governance:-An automatic data mining in distributed environment can create serious issues in terms of data privacy, security and governance. These issues can be addressed by GRID data mining technology.

User-friendliness:-A good System must hide complexity from user. To provide this facility new software, tools and infrastructure is required in the areas of grid-supported workflow management, resource identification, allocation, and scheduling, and user interfaces.

Conclusion:-

As Data Mining is beneficial for government, enterprise and for an individual, helpful in making efficient and effective strategy which can improve their throughput. But there are many challenges and issues in data mining which have to be handled properly such as misuse of information. Data mining provide capability to enterprise/government to extract information, knowledge from huge amount of data and to discover patterns which can help in understanding Customer Behaviour and market trends.

REFERENCES:

[1] Bhoj Raj Sharma, Daljeet Kaur and Manju"
A Review on Data Mining: Its Challenges, Issues and Applications", published in 2013 International Journal of Current Engineering and Technology.

[2] Savasre A., Omienciski E., and Navathe S., (1995), an efficient algorithm for mining association rules in large databases. In the proceeding of 21st international conference on *VLDB*, pp. 432-444

[3] Yanthy W., Sekiya T., Yamaguchi K., (2009), Mining Interesting Rules by Association and Classification Algorithms. In the proceeding of International Conference on Frontier of Computer Science and Technology, pp. 177182.

[4]Kusiak A., Kernstine K.H., Kern J.A., McLaughlin, K.A., and Tseng, T.L. (2000), Data Mining: Medical and Engineering Case Studies. Proceedings of the Industrial Engineering Research Conference, Cleveland, Ohio, pp. 1-7.

[5]Luis R., Redol J., Simoes D., Horta N., Data Warehousing and Data Mining System Applied to E-Learning, Proceedings of the II International Conference on Multimedia and Information & Communication Technologies in Education, Badajoz, Spain, and December 3-6th 2003.

[6]Venkatadri.M, and Dr. Lokanatha C. Reddy"A Review on Data mining from Past to the Future" published in 2011" International Journal of Computer Applications (0975 – 8887)
Volume 15– No.7, February 2011".