PARADIGM SHIFT IN DATA MINING TECHNIQUES AND ITS APPLICATIONS

Pooja Yadav, Assistant Prof. ,
Deptt of CSIT, MJP Rohilkhand University

Hemant Yadav
Associate Prof. ,Deptt of CSIT, PGIE,

Abstract: Because of the rapid advancement of technology in today's society, the amount of data saved has been steadily expanding in every industry. Using data mining techniques, it is hoped to extract relevant, valuable information from these data that was previously unknown. Data mining is a critical phase in the information discovery process in databases, and it is a prominent subfield in knowledge management. In the coming decades, data mining research will continue to flourish in business and learning organizations.

The definition of data mining its associated terminologies are discussed and then data mining techniques are briefly described in the first section of the article. Then, useful data mining tools are described. Finally, data mining applications in the are presented and explored.

Keyword: Web mining, Knowledge Discovery in Databases (KDD), Data warehouse, Data Extraction, Classification, Clustering, Association, Regression

Introduction:

The World Wide Web is a large repository of data and services that is continually expanding. Search engines have been created to assist in the discovery of unfamiliar materials by category, contents, or subject.Search engines use massive indexes of web documents to find the URLs of those documents that satisfy a user's query. Query results are frequently inconsistent, with document referrals that satisfy the search requirements but are irrelevant to the user[1].

Data mining is the process of analyzing massive amounts of data stored in computers. Grocery retailers, for example, collect a lot of data from our purchases. Bar coding has made checkout much more convenient for us, while also providing massive amounts of data to retail organizations. Grocery stores and other retail establishments are able to process our purchases swiftly and properly estimate product prices using computers. These same computers may assist retailers with inventory management by calculating the number of units of each product on hand in real time. Data obtained by bar coding, along with a variety of other types of data, can be used for data mining analyses [2].

Background:

Need of Data Mining:

Data mining necessitates the identification of a problem, the gathering of data that can lead to a better understanding, and the use of computer models to do statistical or other types of analysis. This can be aided by data visualisation tools or by performing basic statistical analysis, such as correlation analysis.

Data mining technologies must be adaptable, scalable, and capable of properly anticipating responses between actions and outcomes. They must also be automated. The capacity of a tool to apply a wide range of models is referred to as versatility. If a tool works on a small data set, it should also operate on bigger data sets, according to scalable tools [2].

Basic Terminology[3]:

*Data Mining:*The technique of extracting information from large quantities of data is known as data mining. In other terms, data mining is the process of extracting knowledge from data.
*Data Warehouse:*A data warehouse is a location where data may be kept for later analysis. It's similar to a fast computer with a massive data store capacity. Data is transferred from various organizations' systems to the Warehouse, where it may be retrieved and checked for faults.
Web Mining: Web mining is the process of using data mining tools to find patterns on the internet. It extracts structured and unstructured data from web pages, server logs, and link structures using automated approaches.
Text Mining: Text mining, also known as text data mining, is the technique of extracting high-quality information from text. It is comparable to text analytics. It entails "the automatic extraction of information from several written resources by a computer to discover new, previously undiscovered information.
Knowledge discovery in databases (KDD): KDD is a multistep process for discovering relevant information and patterns in data, and Data Mining is one of the processes in KDD for extracting patterns using algorithms.
KDD Procedures: It has several steps. In this context, data mining refers to a particular step of this process .They are:

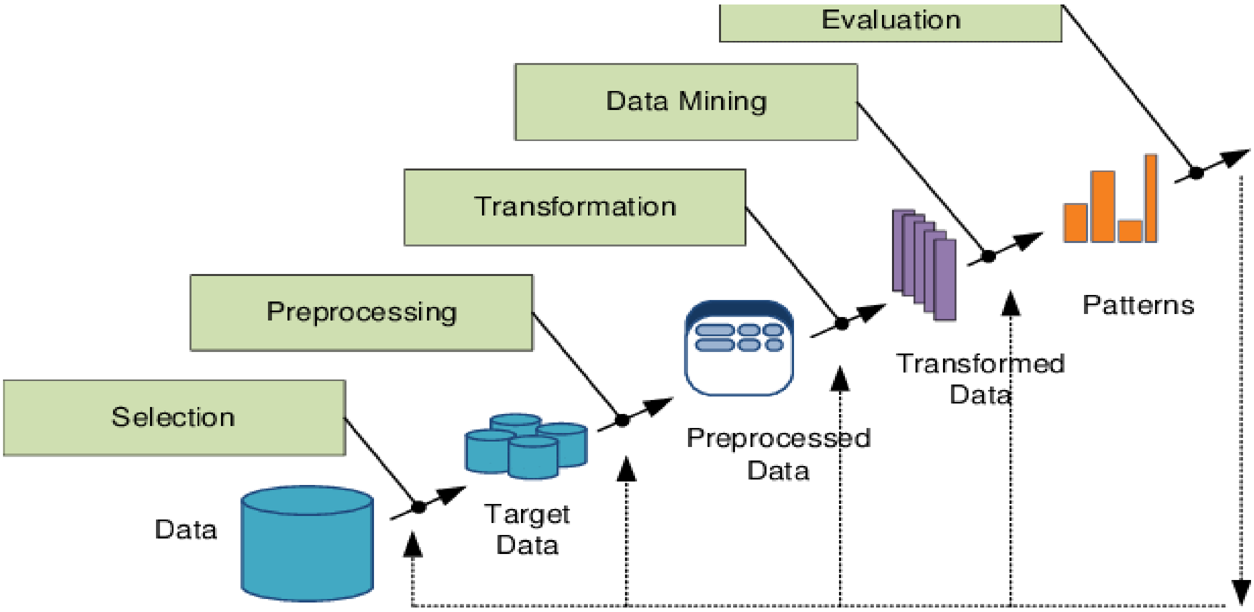

Fig 1: Steps in KDD Process

1. Data Selection/Extraction - Obtaining data from a variety of data sources such as databases, data warehouses, the internet, and other information repositories
2. Data Preprocessing/Cleaning - Incomplete, noisy, and inconsistent data to be cleaned-Missing data may be disregarded or anticipated, and erroneous data may be removed or rectified.
3. Data Transformation / Integration- Combines data from many sources into a logical store-Data can be encoded in standard formats, normalised, and compressed.
4. Data Mining - Use algorithms to extract patterns from altered data.

5. Evaluation/interpretation of patterns- Evaluate the interestingness of the patterns that result, or use interestingness measurements to filter out patterns that are identified.Knowledge visualization methods may be used to show the mined knowledge.

Models of Data Mining:

Mainly two types of data mining models are used [4]:

1. A *predictive model* is one that is built to forecast a specific result or target variable. Multiple regression (for predicting value data), logistic regression (for response prediction), and decision trees (for rule-based value or response models) are all common predictive modeling techniques.

2. *Descriptive model* is a model that improves data understanding without focusing on a particular goal variable. Factor analysis (for extracting underlying dimensions from multivariate data), cluster analysis (for segmenting a client database), and association analysis (for detecting associations between elements such as retail prices) are all common descriptive approaches.

Data MiningTask/Techniques:

There are basicsix main techniques or functions of data mining for which Data mining is in trend are[5]:

1. *Classification:* Finding models to examine and classify a data item into numerous predetermined classifications is referred to as classification.

2. *Regression*: The mapping of a data item to a real-valued prediction variable is known as regression.

3. *Clustering*: Clustering is the process of determining a limited number of categories or clusters that can be used to explain data.

4. *Association:* Finding a model that describes substantial relationships between variables is called dependency modelling (Association Rule Learning).

5. *Anomaly Detection*: Anomaly Detection (Deviation Detection) is the process of identifying the data's most important changes.

6. *Summarization:* Finding a concise explanation for a subset of facts is what summarization is all about.

Data Mining Tools:

Data mining is a powerful new tool that may help firms focus on the most relevant information in their data warehouses by extracting hidden predictive information from massive databases. It discovers and presents knowledge in a way that humans can understand using machine learning, statistical, and visualization techniques. Today, a variety of popular data mining technologies are available.

Data mining software forecasts future trends and behaviours, allowing firms to make informed, proactive decisions. Data mining techniques can provide answers to business questions that were previously too time consuming to answer[4][6][7].

Table 1: Data Mining Tools and their discription

1	Traditional Data Mining Tools	Using a variety of complex algorithms and approaches, traditio applications assist businesses in identifying data patterns and t these programmes are placed on the desktop to monitor data and while others take data that is not stored in a database. Although m in both Windows and UNIX versions, others are only for one o Furthermore, while some may focus on a single database type, t be able to handle any data using online analytical processin technology.
2	Dashboards	A dashboard is a type of graphical user interface that displays k indicators (KPIs) pertinent to a specific goal or business activity other contexts, "dashboard" refers to a "progress report" or "repor of data display.
3	Text-mining Tool	Using a variety of complex algorithms and approaches, traditional data mining applications assist businesses in identifying data patterns and trends. Some of these programmes are placed on the desktop to monitor data and identify trends, while others take data that is not stored in a database. Although most are available in both Windows and UNIX versions, others are only for one operating system. Furthermore, while some may focus on a single database type, the majority will be able to handle any data using online analytical processing or equivalent technology.
4	WEKA tool	Weka (Waikato Environment for Knowledge Analysis) is a Java-based set of machine learning tools. A GUI presentation that contains a variety of visualization tools for predictive modeling, allowing you to construct and test data models while visually analyzing model performance.
5	RAPIDMINER TOOL	Rapid Miner is a popular predictive analytic technology developed by the Rapid Miner corporation. It's programmed in the Java programming language. Text mining and predictive analysis are all possible in this integrated ecosystem.
6	SAS	SAS Institute has developed a statistical software suite called SAS (formerly "Statistical Analysis System"). SAS is a statistical analysis software suite that can explore, edit, organise, and retrieve data from a number of sources. Advanced analytics, multivariate analysis, business intelligence, criminal investigation, and predictive analytics are just a few of the applications. The SAS language includes a graphical point-and-click user interface for non-technical users, as well as other features.
7	TANAGRA TOOL	TANAGRA is a free DATA MINING software that can be used in academic and research settings. In the areas of exploratory data analysis, statistical learning, machine learning, and databases, it presents numerous data mining methodologies.
8	DBMINER TOOL	DBMiner is a data mining technology that allows users to

		interactively mine many levels of knowledge in huge relational databases. It is based on data mining techniques as well as DBLearn, an early system prototype.
9	KNIME TOOL	Using a variety of complex algorithms and approaches, traditional data mining applications assist businesses in identifying data patterns and trends. Some of these programmes are placed on the desktop to monitor data and identify trends, while others take data that is not stored in a database. Although most are available in both Windows and UNIX versions, others are only for one operating system. Furthermore, while some may focus on a single database type, the majority will be able to handle any data using online analytical processing or equivalent technology.
10	ORACLE DATA MINING	Oracle data mining software, which is part of Oracle Advance Analytics, offers excellent data mining algorithms for data classification, prediction, regression, and specialised analytics, allowing analysts to analyse insights, make better predictions, target best customers, find cross-selling opportunities, and detect fraud.

Applications of Data Mining[8][9]**:**

Healthcare: Data mining has a lot of promise for improving health-care systems. It identifies best practices for improving treatment and lowering costs using data and analytics. Multi-dimensional databases, machine learning, soft computing, data visualization, and statistics are among the data mining techniques used by researchers. The volume of patients in each group can be predicted using data mining. Patients receive appropriate care at the correct place and at the right time thanks to the development of processes. Healthcare insurers can employ data mining to detect fraud and misuse.

Market Basket Analysis: Data mining techniques used in Market Basket Analysis. When a consumer wants to buy anything, this technique aids us in determining the relationships between the many goods that the customer has placed in their shopping carts. The finding of such relationships, which enhances the business technique, may be found here. In this approach, retailers employ data mining techniques to determine which customers' intentions are (buying the different pattern). In this way, the strategy is employed to increase business revenues while also assisting in the purchase of connected things.

Retailing: The retail business collects a lot of information on sales and client buying habits. The amount of data collected is continuously expanding, owing to the increasing accessibility, availability, and popularity of conducting business over the internet, or e-commerce. Data mining opportunities abound in the retail industry. Retail data mining can aid in the identification of customer behavior, the discovery of customer shopping patterns and trends, the improvement of customer service, the improvement of customer retention and satisfaction, the improvement of goods consumption ratios, the design of more effective goods transportation and distribution policies, and the reduction of business costs.

Banking and Finance: The banking industry has seen significant transformations in the way it does business around the world. The recording of transactional data has become easier with the recent installation, broader acceptance, and use of "electronic banking," while the volume of such data has expanded significantly. Analyzing this massive volume of raw data and properly transforming the data into meaningful knowledge for the organisation is beyond human capability.

Education: Educational Data Mining is a rapidly growing subject that is concerned with developing ways for discovering knowledge from data originating from educational environments. Predicting students' future learning behaviour, investigating the benefits of educational support, and expanding scientific knowledge about learning are all goals of EDM. An institution can employ data mining to make informed decisions and anticipate student outcomes. As a result of the findings, the institution can concentrate on what to teach and how to teach it. Students' learning patterns can be collected and used to build teaching strategies.

Challenges related to Data mining:

There are some Data Mining risks[4], such as:

(a) Issues with data quality — the data being mined must be of high quality, consistency, and integrity. Failure to do so, both throughout the modeling and deployment stages of the process, can be disastrous.

(b) Results produced by untrained users using highly automated modeling tools can be deceptive or illogical.

(c) Generating masses of useless or non-actionable data. It's only beneficial to be able to spot patterns in a data warehouse if there's a business application. Having a large number of patterns but no revenue-generating applications might be a costly diversion.

(b) Poor model efficacy evaluations or a lack of guidelines for evaluating descriptive data can lead to misapplication of the findings and no gains from the procedure.

(e) In the modeling step, certain technical criteria apply (such as avoiding projecting outside the scope of the data), which is why users must be fully trained.

References:

[1] S. S. Bedi, H. Yadav, and P. Yadav, “Categorization, Clustering and Association Rule Mining on WWW,” in *2009 International Multimedia, Signal Processing and Communication Technologies, IMPACT 2009*, pp. 173–177,2009.

[2] D. L. Olson, *Advanced Data Mining Techniques*. Springer, 2008.

[3] “Data mining - Wikipedia.” https://en.wikipedia.org/wiki/Data_mining

[4] B. Leventhal, “An introduction to data mining and other techniques for advanced analytics,” *J. Direct, Data Digit. Mark. Pract.*, vol. 12, no. 2, pp. 137–153, Oct. 2010.

[5] T. Silwattananusarn and A. KulthidaTuamsuk, “Data Mining and Its Applications for Knowledge Management : A Literature Review from 2007 to 2012,” *Int. J. Data Min. Knowl. Manag. Process*, vol. 2, no. 5, 2012.

[6] Y. Ramamohan, K. Vasantharao, C. K. Chakravarti, and A. S. K. Ratnam, “A Study of Data Mining Tools in Knowledge Discovery Process,” *Int. J. Soft Comput. Eng.*, vol. 2, no. 3, pp. 191–194, 2012, ISSN: 2231-2307.

[7] J. Silltow, “Data Mining 101: Tools and Techniques,” 2006.

[8] P. Perner, Ed., “Advances in Data Mining: Applications and Theoretical Aspects,” 2010.

[9] N. Padhy, P. Mishra, and R. Panigrahi, “The Survey of Data Mining Applications And Feature Scope Keywords Data mining task, Data mining life cycle , Visualization of the data mining model , Data mining Methods, Data mining applications,” *Int. J. Comput. Sci. Eng. Inf. Technol.*, vol. 2, no. 3, 2012.