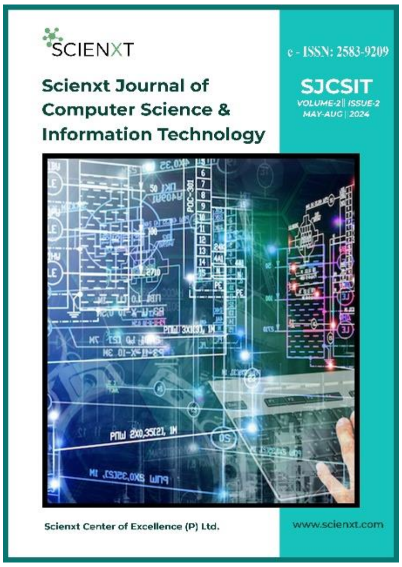

Scienxt Journal of Computer Science & Information Technology
Volume-2 || Issue-2 || May-Aug || Year-2024 || pp. 1-13

A review on big data and data mining concepts

Priyanka Singh
Assistant Professor, CSE Department BITS, Bhopal, Madhya Pradesh, India

**Corresponding Author: Priyanka Singh*
Email: priyankasinghcse02@gmail.com

Abstract:

Abstract - In the digital era like today the growth of data in the database is very rapid, all things related to technology have a large contribution to data growth as well as social media, financial technology and scientific data. Therefore, topics such as big data and data mining are topics that are often discussed. Data mining is a method of extracting information through from big data to produce an information pattern or data anomaly. In this paper discussing the data mining technology, its models, task, applications, comparision between big data and data mining.

Keywords:

Data mining, big data, Data mining applications.

1. Introduction:

Big Data: It is huge, large or voluminous data, information or the relevant statistics acquired by the large organizations and ventures. Many software and data storage created and prepared as it is difficult to compute the big data manually. It is used to discover patterns and trends and make decisions related to human behavior and interaction technology. Big data comprises of 5Vs that is Volume, Variety, Velocity, Veracity, and Value.

Volume: In Big Data, volume refers to an amount of data that can be huge when it comes to big data.

Variety: In Big Data, variety refers to various types of data such as web server logs, social media data, and company data.

Velocity: In Big Data, velocity refers to how data is growing with respect to time. In general, data is increasing exponentially at a very fast rate.

Veracity: Big Data Veracity refers to the uncertainty of data.

Value: In Big Data, value refers to the data which we are storing, and processing is valuable or not and how we are getting the advantage of these huge data sets.

Data Mining: Data Mining is a technique to extract important and vital information and knowledge from a huge set/libraries of data. It derives insight by carefully extracting, reviewing, and processing the huge data to find out pattern and co- relations which can be important for the business. It is analogous to the gold mining where golds are extracted from rocks and sands.

2. Big data vs data mining:

Table. 1: Differences between Big data and Data mining

Data Mining	*Big Data*
It primarily targets an analysis of data to extract useful information.	It primarily targets the data relationship.
It can be used for large volume as well as low volume data.	It contains a huge volume of data.
It is a method primarily used for data analysis.	It is a whole concept than a brief term.

It is primarily based on Statistical Analysis, generally target prediction, and finding business factors on a small scale.	It is primarily based on data analysis, generally target prediction, and finding business factors on a large scale.
It uses the following data types e.g., Structured data, relational, and dimensional database.	It uses the following data types e.g., Structured, Semi- Structured, and unstructured data.
It expresses what about the data.	It refers to why of the data.
It is the closest view of the data.	It is a broad view of the data.
It is primarily used for strategic decision- making purposes.	It is primarily used for Dashboards and predictive measures.

3. Literature survey:

According to Connolly et al. 1999 [1] Data mining is “a process of extracting valid, previously unknown, understandable, and actionable information from huge databases and using it to make essential business decisions”.

Rygielski et al. 2002 [2] describe the relationship marketing a reality. Technologies such as data warehousing, data mining and operations management software have prepared customer relationship management a new area where firms can gain a competitive advantage. Particularly through data mining the extraction of unknown predictive information from huge databases organizations can identify valuable customers, predict future behaviors, and permit firms to make proactive, knowledge-driven decisions.

Yin et al. 2004 [3] study, the characteristics of the FEA data are discussed firstly. Then a framework of knowledge discovery from FEA data is proposed. In the same way, a data-mining algorithm named fuzzy-rough algorithm is developed to deal with the FEA simulation data. Finally, the stamping process of a square-cup part was an example. The proposed knowledge discovery process is applied to obtain some useful, understood production rule with efficiency measure.

According to Alhammdy et al. 2007 [4] Streaming data mining is one of the most difficult tasks in Knowledge Discovery in Databases (KDD). In this paper, study the meaning of emerging patterns in data streams by introducing a special type of emerging patterns, matching the emerging pattern (MEPs). This type of EPs can be easily mined from data streams by applying a selective approach to conduct the mining process. This experiment proves that MEPs are capable of gaining important information from streaming data. This information increases the accuracy of classification.

Liu et al. 2010 [5] presents the technology of the process knowledge discovery in the process database. After analyzing the process planning knowledge discovery flow and its key technologies are also discussed. It has many advantages. Furthermore, it can accelerate the standardization of process planning. Finally, the PPK discovery system is designed and the structure and function of the system are stated.

Diamantini et al. 2011 [6] introduces Designer, a web based semantic driven tool intended at supporting users in the mutual design of a KDD process. A designer, a tool for supporting non-expert users in the mutual design of KDD processes. By exploiting an SOA-based methodology, execute KDD tools as web services, solving the heterogeneity of their interfaces, and allowing a typical communication protocol.

Chen Zhang et. al. 2023 [7] this paper aims at metro station clustering based on passenger flow data. Compared with existing clustering methods that only use boarding or alighting data of each station separately, we focus on higher granularity origin-destination (O-D) path flow data, and provide more flexible and insightful clustering results. Qiyi He et. al. 2023 [8] in this paper, a novel hybrid ARM method based on WWO with Levy flight (LWWO) is proposed. The proposed method improves the solution of WWO by expanding the search space through Levy flight while effectively increasing the search speed. In addition, this paper employs the hybrid strategy to enhance the diversity of the population in order to obtain the global optimal solution. Moreover, the proposed ARM method does not generate frequent items, unlike traditional algorithms (e.g., Apriori), thus reducing the computational overhead and saving memory space, which increases its applicability in real-world business cases.

To review, data mining is a way to find previously unknown, valid patterns and relationships from the huge amount of data represented in qualitative, textual, or multimedia formats by applying different data analysis tools and also most of the time the datasets are collected for other purposes.

4. Architecture and process of data mining:

4.1. Architecture of data mining:

Data mining is the process of discovering interesting knowledge of the huge amount of data stored in the data warehouse, databases or other information repositories. Based on this analysis, the architecture of a typical system has the following major components as shown in fig. 1:

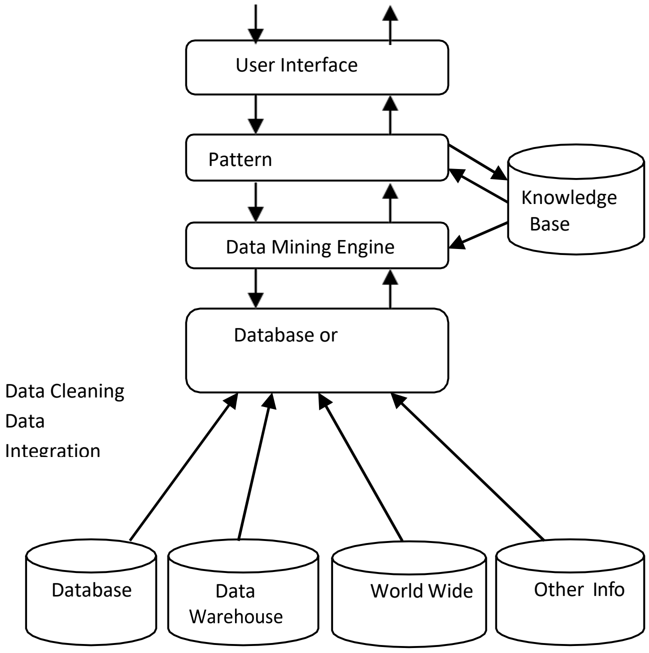

Figure. 1: Architecture of typical data mining system

1) Data warehouse, database, World Wide Web, or other information repository: This is one or the set of the data warehouse, databases, spreadsheets, or other kind of information repositories. Data cleaning & data integration techniques may be performing of the data.

2) Database or data warehouse server: -This is responsible for fetching the relevant data, based on the user's data mining request.

3) Knowledge base: This is the domain knowledge that is used to guide the search or analyzes the interestingness of the resulting pattern. Such knowledge can include the concept hierarchy & user viewpoint.

4) Data mining engine: This ideally important to the data mining system & consists of sets of functional component of tasks such as characterization, association & correlation analysis, classification, prediction, cluster analysis, outlier analysis & evolution analysis.

5) Pattern evaluation module: This component that usually includes interestingness measures & interacts with the data mining module so as to focus the search towards interesting pattern. The pattern estimate method can be integrated with data mining component depending on the implementation technique used.

6) User interface: This module converse between the user & the data mining system, allow the user to interact with the system by specifying a data mining query or task, given that information to help focus the search & performing the tentative data mining based on the transitional data mining results.

4.2. Process of data mining:

The KDD process is interactive and iterative, involving numerous steps with many decisions being made by the user. Each step attempts to complete a particular discovery task and each accomplished by the application of a discovery method. Knowledge discovery concerns the entire knowledge extraction process, including how data are stored and accessed, how to use efficient and scalable algorithms to analyze massive datasets, how to interpret and visualize the results, and how to model and support the interaction between human and machine. It also concerns support for learning and analyzing the application domain.

Many people treat the data mining as a synonym for generally used term, Knowledge Discovery from Data. Others analysis the data mining as simply a crucial step in the process of knowledge discovers as shown in Fig. 2.

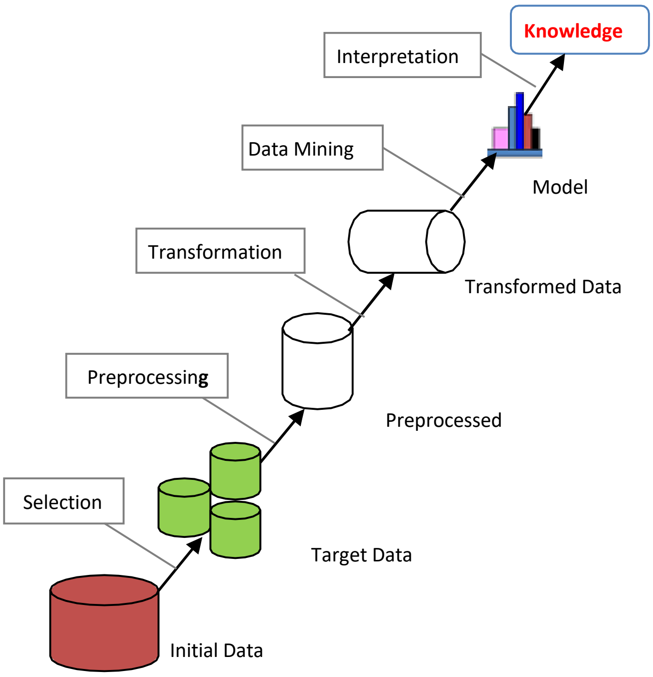

Figure. 2: Data Mining Process

1) Data selection: Selecting the data required for data mining process & may be obtained from many different & various data sources.

2) Data preprocessing: This includes result incorrect or missing data. There may be several different activities performed at this time. Flawed data may be corrected or removed,

whereas missing data must be supplied. Preprocessing also includes: removal of noise or outliers, collecting essential information to model or account for the noise, accounting for time sequence information and known changes.

3) Data transformation: This converting the data into a common format for processing. Some data may be encoded or transformed into a more functional format. Data reduction, dimensionality reduction (e.g. Feature selection i.e. Attribute subset selection, heuristic method etc.) & data transformation method (e.g. Sampling, aggregation, generalization etc) may be used to reduce the number of possible data values being measured.

4) Data mining: An important process where intellectual techniques are applied to orders to mine data patterns.

5) Interpretation/evaluation: To identify how the data mining results are obtainable the users which are extremely important because the utility of the result is dependent on it. A variety of visualization & GUI strategies are used in this step. A different kind of knowledge requires different kinds of representation, e.g. Clustering, classification, association rule etc.

5. Data mining tasks:

Data mining tasks are used to classify the kind of patterns to be created in the data mining process. In general, data mining tasks can be classier into two categories: Predictive and Descriptive. A Predictive model makes a prediction about values of data using well-known results found from different data and its objective is to discover strong links between variables of a data table (columns). A descriptive model classifies patterns or relationships in data. It simply summarizes data in suitable behavior or in ways that will lead to improved considerate of the way things work. The major difference between the two models is that, a descriptive model serves as a way to discover the properties of the data examined, not to predict new properties. In contrast, a predictive model has the specific goal of allowing us to predict the value of some target typical of an object on the basis of the practical values of other distinctiveness of the object.

Predictive model data mining tasks contain classification, prediction, regression, and time series analysis. The Descriptive task encompasses methods such as Clustering, Summarizations, Association Rule Discovery, and Sequence analysis.

5.1. Classification:

Classification is that the method that finds the common properties among a group of objects in a database and classifies them into totally different classes, consistent with a classification model. The objective of the classification is to first analyze the training data and develop an accurate description or a model for every class using the options available within the data such class description are then used to classify future test data. Such class descriptions are then used to classify future test data within the database or to develop an improved description for every class within the database. Some common classification strategies incorporate, support vector machines, decision trees, and logistic regression.

5.2. Prediction:

There are two main varieties of predictions: one will either attempt to predict some occupied data values or during lean, or predict a class label for only some data and is tied to classification. Once a classification model is completed to support a training set, the class label of an object will be foreseen supported the feature values of the object and also the characteristic values of the classes. Prediction is observed the forecast of missing numerical values, or increase/ decrease leaning in time related data. The mainly significant idea is to use a large range of past values to treat as potential future values.

5.3. Regression:

Regression technique also can be adapted for prediction. In regression, the predicted variable may be a continuous variable. The regression involves the learning of function that map data item to a true valued prediction variable. Some common regression strategies include statistical regression, neural networks and support vector machine regression. Several real-world data mining issues don't seem to be merely predictive. So more complex techniques may be necessary to forecast future values using a combination of the techniques (e.g. logistic regression, decision trees or neural networks).

5.4. Time series analysis:

In the time series analysis the value of an attribute is examined as it varies over time. In time series analysis is used for many statistical techniques which will analyze the time- series data such as auto regression methods etc. It is sometimes used in the two types of modeling (i) ARIMA (ii) Long-memory time-series modeling.

5.5. Clustering:

The process of grouping physical or abstract objects into classes of similar objects is called clustering or unsupervised classification. Clustering constitutes a major class of data mining and a standard technique for statistical data analysis used in many fields; involve pattern recognition, info retrieval, machine learning, Bioinformatics, and image analysis. Cluster analysis itself isn't one specific algorithmic rule, but the ultimate task to be solved. It's usually achieved by completely different type algorithms that produces an effort to automatically partition the data space into a group of regions or clusters, to that the examples within the table are assigned, either deterministically or probability wise. The aim of the method is to identify all set of similar examples within the data, in some optimal fashion.

5.6. Summarization:

Summarization, also referred to as Description or Generalization, pulls the data into subsets with their various descriptions. Generally actual parts of the mined data are retrieved and supported that the subsets described. Summarization isn't a data Mining method; it's the result of data Mining technique.

5.7. Association rule mining:

Association rule mining discovers relationships among attributes within the dataset, manufacturing if-then statements regarding attribute-values [9]. Association rule mining is one among the necessary technique that aims at extracting, interesting correlations, frequent patterns, associations or casual structures among set of items within the transaction databases. An X => Y association rule expresses a close relationship between items (attribute-value) during a database with values of support and confidence. Association analysis is usually used for market basket analysis [10].

5.8. Sequence discovery:

Sequence discovery is used to see sequential patterns within the data. These sequences are more typically associations between variable data fields, however they're primarily based on time and sometimes follow a specific queue. This method encompasses association rules similarly as Markov concepts; hence not much can be elaborate on concerning this. As an example, if someone gets an electronic equipment then he's certain to buy CDs for it earlier than later.

6. Applications of data mining some applications of data mining are:

6.1. Data mining applications in healthcare:

Data mining applications can significantly advantage all parties engaged in the healthcare [11] industry. For example, data mining can facilitate healthcare insurers detect fraud and abuse health care organizations make customer relationship management decisions, physicians identify effective treatments and best practices, and patients get better and more affordable healthcare services.

The enormous amounts of data produced by healthcare transactions are also complex and huge to be processed and analyzed by traditional methods. Data mining provides the methodology and technology to transform these mass of data into useful information for decision making.

6.2. Educational data mining:

At present there is an increasing interest in data mining and educational systems, making educational data mining as a novel rising research society. The application of data mining to conventional educational systems, mostly web-based courses, illustrious learning satisfied management systems, and adaptive and intelligent web-based educational systems [12]. Each of these systems has a dissimilar data source and purpose for knowledge discovering. After preprocessing the accessible data in each case, data mining techniques can be applied: statistics and visualization; clustering, classification, and outlier detection; association rule mining and pattern mining; and text mining.

Educational data mining [13] is an emerging trend, concerned with developing techniques for exploring, and analyzing the huge data that come from the educational context. EDM is poised to leverage an enormous amount of research from the data mining community and apply that research to educational problems in learning, cognition, and assessment. In recent years, Educational data mining has proven to be more successful at many of these educational statistics problems due to enormous computing power and data mining algorithms.

6.3. E-commerce is also the most prospective:

Electronic commerce (EC) [14] has become a trend in the world nowadays. However, most researches neglect a fundamental issue – the user’s product-specific knowledge on which the useful intelligent systems are based. This research employs the user’s product-specific knowledge and mine his/her interior desire for appropriate target products as a part of the personalization process to construct the overall EC strategy for businesses.

In order to facilitate transactions, the problems associated with complex activities in electronic commerce must be resolved. The abundance of information available on the Internet allows consumers to communicate with sellers for a bargain. Therefore, the traditional commerce negotiation process, similar to human-based life bargaining between buyers and sellers, will also arise in the electronic market in order for both parties to reach an agreement that is satisfactory to both.

6.4. Sports data mining:

The sports [15] world is known for the vast amounts of statistics that are collected from each player, team, game, and season. There are also many types of statistics that are gathered for each – a basketball player will have data for points, rebounds, assists, steals, blocks, turnovers, etc. for each game. This can result in information overload for those trying to derive meaning from the statistics. Hence, sports are ideal for data mining tools and techniques.

6.5. Data mining is used for market basket analysis:

Data mining technique is used in MBA (Market Basket Analysis) [16]. When the customer wants to buy some products then this technique helps us finding the associations between different items that the customer puts in their shopping pockets. Here the discovery of such associations that promotes the business technique .In this way the retailers use the data mining technique so that they can identify that which customers intension (buying the different pattern). In this way this technique is used for profits of the business and also helps to purchase the related items.

7. Conclusion:

The data mining techniques can be applied on big data to acquire some useful information from large datasets. Thus these two terms are not different instead they are coupled together to acquire some useful picture from the data. Thus we conclude that big data will become an excellent opportunity in the fourth coming years. We discussed some of the useful information about big data and data mining and have identified the research gaps and open research areas.

8. References:

(1) T. Connolly, C. Begg, and A. Strachan "Database Systems: A Practical Approach to Design, Implementation and Management," Second Edition. Addison-Wesley, New York 1999.

(2) C. Rygielski, J.C. Wang, D. C. Yen, "Data mining techniques for customer relationship management," Technology in Society 24, 2002, pp. 483–502.

(3) J.L.Yin, D.Y. Li, Y.C. Wang, Y.H. Peng, "Knowledge Discovery from Finite Element

Simulation Data," IEEE Proceedings of the Third International Conference on Machine, Learning and Cybernetics, Shanghai, August 2004, pp. 26-29.

(4) H. Alhammady, "A Novel Approach for Mining Emerging Patterns In Data Streams," IEEE 2007.

(5) S. Liu, X. Tian, Z. Zhang, "Process Planning Knowledge Discovery in the Process Database," IEEE International Conference on Computer Application and System Modeling (ICCASM) 2010, pp. 370-373.

(6) C. Diamantini, D. Potena and E. Storti, "A Semantic-Aided Designer for Knowledge Discovery," IEEE 2011, pp. 86-93.

(7) Zhang, C., Zheng, B. & Tsung, F. Multi-view metro station clustering based on passenger flows: a functional data-edged network community detection approach. Data Min Knowl Disc (2023).

(8) He, Q.; Tu, J.; Ye, Z.; Wang, M.; Cao, Y.; Zhou, X.; Bai, W. Association Rule Mining through Combining Hybrid Water Wave Optimization Algorithm with Levy Flight. Mathematics 2023.

(9) R. Agarwal, T. Imielinski, & A. Swami, "Mining association rules between sets of items in large databases," In Proceedings of the ACM SIGMOD international conference on management of data, Washington DC, USA, 1993, pp. 1–22.

(10) Z. Zheng, R. Kohavi, L. Mason, "Real World Performance of Association Rule Algorithms," Proceedings of the Seventh ACM- SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, ACM, 2001.

(11) H. C. Koh and G. Tan, "Data Mining Applications in Healthcare," Journal of Healthcare Information Management Vol. 19, No. 2, pp. 64- 72.

(12) C. Romero, S. Ventura, "Educational data mining: A survey from 1995 to 2005," Expert Systems with Applications 33, 2007, pp. 135–146.

(13) S.R Barahate, V.M. Shelake, "A Survey and Future Vision of Data mining in Educational Field," IEEE Second International Conference on Advanced Computing & Communication Technologies 2012, pp. 96- 100.

(14) S. Ansari, R. Kohavi, L. Mason, and Z. Zheng, "Integrating E- Commerce and Data Mining: Architecture and Challenges," Proceedings of IEEE International Conference on Data Mining, 2001.

(15) O.K. Solieman, "Data Mining in Sports: A Research Overview," A Technical Report, MIS Master Project, August 2006, pp. 1-76.

(16) T. Raeder, N. V. Chawla, "Market Basket Analysis with Networks," Springer 2011.