Journal of Advances and Scholarly Researches in Allied Education
Vol. 16, Issue No. 2, February-2019, ISSN 2230-7540

Review on Data Mining

Urvashi Sangwan*

Assistant Professor, Department of Computer Science and Engineering, Vaish College of Engg, Rohtak

Abstract – Data, information and knowledge are the interesting role of human life. Huge repositories of data with the fast development of technologies have required analyzing and Modeling of big data to predict and analyze the future trends of information. Knowledge discovery in the databases needs methodologies and techniques used into various areas of information systems. Data mining is a knowledge discovery that extracts useful information. It has been a major advance in machine learning, artificial agent systems, and decision making in the expert systems. The last decade, the researcher has surveyed most of the techniques and applications that used in different fields in our life such as manufacturing, education, engineering and business.

Keywords: Data Mining Process, Knowledge Discovery, Database, Techniques

----------X----------

I. INTRODUCTION

Every large data set maintains a pattern which may not be understandable all of a sudden. Data mining is a process of discovering the patterns in the large data set. The aim of data mining is to extract information from large datasets and modify them to usable structures so that the extracted information can be applied in proper place and time without difficulty. This is controlled through databases with different database management aspects included. Data mining is a commonly used term in any kinds of data processing these days. Data scientists started using the term in the early 1990s. Earlier it was also referred by some other names such as Knowledge Discovery in Databases (KDD), Data Science, and Productive Analytics [1]. Data mining is a general interactive and iterative discovery process. Data mining has multiple applications such as determining mine patterns, data fields association, changes in data, anomalies in a data set, and determining statistically viable data structures [3]. The mined results should be valid, novel, applicable, and recognizable.

A. Objective of the study

The objective of the study is to find the reviews on the Data Mining with Techniques.

II. PROCESS OF DATA MINING

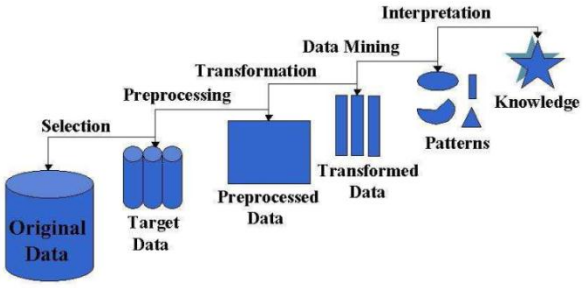

Figure 1: Data mining process **[3]**

1. Data hauling, transforming and loading transaction data in the data warehouse system.
2. Storing and managing data in the multidimensional database system.
3. Make provision for data access to experts in the fields of business analysis and information technology.
4. Use of software for data analysis.
5. Data presentation through easily understandable formats such as table and graph.

III. TECHNIQUES OF DATA MINING

Data mining is a multidimensional and multilevel complex process. It requires efficient techniques

and first processing devices. Following are some important techniques in this realm:

A. Designed neural network

This takes the help of artificial intelligence (AI). Techniques like pattern recognition, neural networks, and machine learning are very useful in data mining. Several other techniques like knowledge acquisition and representation, and data searching are relevant to multiple process and steps in data mining [4]. It is a form of non-linear predictive model. It resembles almost like biological neural networks. It learns through training.

B. Genetic Algorithms

There are optimization techniques or processes that take help of processes like a genetic mutation, genetic combination, and natural selection. These techniques are based on the concept of natural evaluation. There are rules that represent a possible solution to a problem is originally created at random. Then, the rules are combined to reproduce next-generation rules. A mutation process is applied to modify randomly the genetic structures of some members of each new generation. This is a continuous process that ends when a comprehensive solution is obtained. As such, genetic algorithms are perfect problems that summons optimization with respect to some computable measures. This concept can be applied in data mining also. If the problems are large and multidimensional, then fast processors are required to obtain appropriate solution within a reasonable amount of time. These days, high-speed computer processors are available that makes data mining rather easier and faster [4].

C. Decision tree

The chain of decisions is given a tree-shaped structure for better understanding. This is commonly known as a decision tree. These decisions develop rules for dataset classification. There are some prominent decision tree processes such as Chi-Square Automatic Integration Detection (CHAID) and Classification and Regression Trees (CART). These techniques are used for the classification of datasets. They offer a set of rules that can be used for the dataset classifications and can also be applied in a new dataset environment to predict which records could provide a given outcome [4].

D. Nearest neighbour method

It is also called K-nearest data technique. This is a technique to classify each record in a dataset on the basis of the combination of the K record(s) classes that are similar to historical dataset (Wherek1).

E. Rule induction

This called “if-then” rule that is based on the techniques of statistical significance.

F. Data visualization

It is related to the visual interpretation of multilevel relationships in a multidimensional data structure. Here, graphic tools are very useful to set relationships among a wide array of data. An image is much useful than a table of numbers. Visual data mining processes are found to be extremely useful in exploratory data analysis and they also have the ability to mine large database. This approach is appropriate for the integration of human with the data mining processes.

IV. KNOWLEDGE DISCOVERY IN DATABASES

The other name of Data Mining, as stated earlier, is Knowledge Discovery in Databases (KDD). It refers to not-so-important extraction if implicit, comparatively less known and useful information out of the datasets or data in a database. Though these two terms are used alternatively in many places, in reality, KDD is a bigger functionality and data mining is a part of it [4]. KDD comprises two major steps that turn raw data to some new forms of knowledge. It consists of the following steps [5]:

Step 1: Data cleaning

This is also called data cleansing. There are several forms of noise data and irrelevant data that have no job in the process. Those datasets are removed in this step.

Step 2: Data integration

At this step, several data sources, often heterogeneous might be combined to form a common source.

Step 3: Data Selection

At this step, relevant data for the process are recognized and retrieved from the collection of data.

Step 4: Data Transformation

At this step, datasets are consolidated and selected datasets are given relevant forms appropriate for data mining.

Step 5: Data mining

This is the most crucial step when different techniques are applied to extract patterns potentially useful.

Step 6: pattern Evaluation

In this step highly interesting patterns that actually forms a new knowledge are identified based on the given parameters.

Step 7: Knowledge representation

In this phase, the newly discovered knowledge is visually presented to the actual users. Here, different visualization techniques are used to make data mining results visible and understandable to the end-users.

V. ISSUES IN DATA MINING

Several data mining algorithm techniques are available that have been existing for years but have been lately been applied as reliable and scalable tools for gathering new knowledge out of datasets. Many of these algorithmic techniques have virtually outperformed older classical tools. Though data mining is still growing in leaps and bounds, it has become an omnipresent tool for many scientific and business jobs, especially where analysis of data matters the most. The days are not far behind when data mining will become a trusted discipline in many areas but before those several aspects of it need to be cleared. Some of these issues are discussed below [2]:

Data Integrity

Quality of data analysis and its relevance depend a lot on the quality of data being analyzed. The key challenge is to integrate different qualities of datasets some of which may be redundant and conflicting from different sources. For example, a bank maintains credit card accounts on several different databases, the problem may occur if the same account holder gives different addresses in different occasions; in that case, the system needs to recognize the account holder and integrate the datasets on the basis of the most recent address entered by the account holder.

Security and Social issues

Security and safety of data in any data collection process is a major issue these days. Sharing of data needs to be hundred percent guarded against all threats. Data mining helps to analyze day-to-day business processes, people's buying habits, and many related matters through gleaning a significant part of information of people's transaction with an organization.

Mining Methodology

A key technical issue is whether relational database structure is better or a multidimensional one? In a relational structure, it is customary to store data in tabular forms and then permitting ad hoc queries. In a multidimensional structure, sets of cubes are formed and then these sets are kept in arrays with subsets are created as per the need of categories. Multidimensional structures are helpful in multidimensional data mining and relational are simple to use. Both of these methodologies have their respective utilities. The internet has made the world a big, uniform, client or server environment.

Issues related to data sources

Sources of data face multiple issues. Some such issues are practical in nature such as data diversity and others are somewhat philosophical in nature like excess data issues.

Cost

Cost of data mining is another big issue. Over the last few years, the cost of data mining hardware systems has dropped quite a lot. As the queries in data mining get powerful, the better and higher the usefulness of information being collected and maintained. This increases the necessity of larges and faster hardware and software systems which are more expensive [5].

VI. REVIEW ON DATA MINING

According to Mostafa (2016), data, knowledge, and information have interesting roles in human life. Availability of huge data, availability of gigantic storing spaces and processes, and invention of really fast data processing techniques have helped to expand the concept of big data and analyze the future trends of information and knowledge. There are several areas of information technology where the discovery of knowledge, related methodologies, and related techniques are applied. Data mining is nothing but discovery of new knowledge or add-on existing knowledge base that ultimately helps to extract useful information. Big data analysis is a major machine learning process, artificial agent system, and decision making in the expert systems [5].

Throughout the last two decades, the researchers have surveyed most of the techniques and applications applied in different fields of human life such as education, business, health, manufacturing, engineering, and agriculture. In this article, most-applied data mining techniques and trends in data mining in the last five years are reviewed closely. While researching and reviewing these techniques and trends relevant in different fields of human life, it is found that in improving the quality of teaching and manufacturing operations, text mining is a process

relevant for getting resourceful information from the data obtained in these sectors.

According to Ghuman (2014), information technology has virtually changed the communication methods replacing traditional, time-consuming, and costly communication processes with cheaper and fast processes. These communication processes take place through several types of modern devices. These devices generate and stores lots of data that require certain purposeful processing. The database technologies are looking for more authentic ways to store, maneuver, and retrieve data while data mining experts are looking for newer, authentic, and faster ways of information extraction from available databases. Data mining is also called Data Science, Predictive Analysis, and Knowledge Discovery in Data Bases (KDD). Today, various techniques are used for artificial neural networks, decision trees, genetic algorithms, visualization, and induction. Data mining is an interactive and iterative discovery process. The goal of this process is to get patters, associations, anomalies, changes and alternations, and statistically relevant data structures out of databases [6].

According to Gera and Goel (2015), data mining is a process of extracting useful data, patterns inherent in a dataset, and data trends from large databases with the help of the techniques such as data clustering, data classification, data association, and data regression. There are different applications of information gleaned from data mining. Different tools are available supporting different algorithms. This paper makes a summary of different data mining tools and supporting respective algorithms. Comparison of different tools has also been accomplished enabling users to apply the tools as per their requirements and applications. In this paper, different validation indices for validation purpose are also summarized [7].

According to Mirza, Mittal, and Zaman (2016), enormous progress in information technology over the last five years has made available huge data in digital form. With the availability of huge data, a new set of challenges related to data retrieval as per the requirements and meaningful information have become very open. Data mining is a tool to tackle this challenge meaningfully. Data mining is considered a stepping stone is knowledge discovery through a rational process and excavate hidden information from databases. Data mining is no accepted as an inherent part of every field of human civilization. In this paper, detail analysis of available literature related to data mining is made, the concept of data mining and related methodologies are summarized, and some tasks, challenges, and applications have been illustrated [8].

According to Manjarres, Sandova & Suárez (2018) data mining in the field of education is an emerging discipline. This seeks to develop methods to explore different types of data from educational institutes. These are related to learning and teaching processes, teacher and student's behaviour, and outcomes of learning processes and effects on students. In recent times, intensive data mining projects have been taken in different educational genres to address educational issues. This paper presents a review of various data mining works and outcomes in education and diverse scenarios in which these data mining techniques have been applied [9].

According to Adebayo and Chaubey (2019), the purpose of data mining is to excavate the patterns included in the datasets and apply them technically to predict a future event. In various educational institutes like colleges, universities, polytechnics, and high schools data mining have been found to immensely useful for the classification and prediction of various outcomes related to various events. Classification is considered as one of the most important techniques for categorizing a specific group of items. The aim of classification is to envisage the character of an item or dataset on the basis of the available classes of items. The classification model is constructed on the basis of available data set. There are different classes of algorithms available in data mining for the assessment of future trends of markets or other business activities. In this paper, the decision tree classification model is used through KNIME tool for assessing the performances of the students in a high school quiz competition [10].

According to Rao, Ramana, and Ramkrishna (2019), image classification has been getting intensive attention in recent times. The authors also opine that with the increasing attention on image classification, the challenges before the use of image classification is also high. The authors speak about a system to reveal the information of the image on the basis of its category and related labels connected with it. This is accomplished through a number of data mining processes and Bag of Visual Words (BoVW) feature for extraction algorithms. Grey level characteristics are applied to develop this algorithm along with some other relevant colour features. Grey characteristics include SURF, i.e. Speed up Robust Features, MESR, i.e. Maximum Stable External Regions, and ICCV, i.e. Improved Colour Coherence Vector. Apart from these techniques, data mining also includes Neural Networks, Bayesian Networks, Decision Trees, K-Nearest Neighbour (KNN), and Discriminant Analysis. These various techniques of data mining are applied in COIL-100, COREL-1000, and many other datasets. It is found that Bayesian Networks and Discriminant Analysis are the best options as they fetch near 100% accuracy, sensitivity, and specificity in case of CORAL-1000 and COIL-100. In later case,

specificity is 100%, accuracy is 98.9%, and sensitivity is 98.5% [11].

According to Maksood and Achuthan (2016), the exponential increase in data volume in recent years has created an urgency is designing log, process, and examine or analyze the vast array of records. Data repositories need to be used scientifically. Bulk irrelevant and unprocessed data require proper management for diminishing the wastage of storage space. Since the early 1990s, various efforts are adopted to redefine and refine the concept of knowledge discovery in data mining and database management systems. Organizations in different industries have been incorporating this approach for predicting buyer behaviour and planning sales promotions. This paper aims to introduce data mining with reviews of real-world applications related to the very basic concept, big data, and data mining techniques. The paper also discusses studies related to smart cities in the matters of energy requirement forecasting and traffic challenges in Oman [12].

According to Salem, Sayed-Mouchaweh, and Hassine (2017), some efficient and relevant criteria are required for different machine learning and data mining methods applied in the field of Residential Energy Smart Management (RESM). The authors proposed a classification to highlight the advantages and restrictions of each category. In this research paper, we find and point out the principal challenges that RESM still faces in different fields [13].

According to Tummala and Kalluri (2018), this is the era of information flow and information sharing. In any case, the general processes of information assessment cannot haul a huge volume of data. The point of focus right now is the way to design an all-encompassing data mining technique to accurately locate all required information swiftly and accurately from volumes of data. In this paper, we start with a precise prologue related to information inquiry and examination followed by the exchanges of huge information examination.

According to Deshpande and Thakare (2010), IT plays a vital role in every sphere of human life. Gathering information from different but reliable data sources and managing them in proper ways by sorting out irrelevant and redundant data is important. As the use of computers and electronic devices are increasing day-by-day, there is an enormous growth in data collection noticed in recent times. Data warehouse needs to be clean and well-managed so that data relevant in a situation could be gleaned anytime. All these activities and related other activities require special tools called data mining tools. In this paper, data mining systems and some of its applications are reviewed [14].

According to Lu, Kim, Zheng, and Jin (2018), the application of big data in the life-cycle of electronic products are pervasive. A complete analysis of big data with DM and review of its different applications at different stages of its application can help to design more reliable and faster-searching tools. In this paper, a clarification is made of DM-related topics. A DM flow chart is given along with main steps involved in the flowchart that is commonly used in the preparation and processing approaches, DM functions and techniques, and key performance indicators. In this paper, we also make a comprehensive review of 105 articles from 2007 to 2017 on DM and big data related to electronics industry to analyze the use of flowcharts from various aspects such as data handling, applications of DM, use of big data in different stages of electronic lifecycles, and software used in the analysis of applications. With the help of all these activities, a related diagram including big data application and DM are established.

According to Sharma, Sharma, and Dwivedi (2017), data mining is a process of data extraction from voluminous datasets. It is an intensive innovative technology. The ultimate goal of data mining process is to get information from a dataset and then modify it as per the requirements. Websites generate millions of unprocessed data and accumulate huge data almost every day. Analyzing this data helps to gain new knowledge. Data mining is not applicable or so effective for unstructured, traditional data [15].

According to Silva and Fonseca (2017), over the last decade, the learning management system has become quite useful in education. Various data mining techniques such as forecasting, clustering, preserving, and relationship building can be applied to educational data. These different applications are useful in determining students' behaviours in different environment and learning outcomes. In this paper, we explore a wide array of data mining techniques that can be applied in educational field. This paper also investigates various recent uses of big data technologies in education and undertakes a detailed literature review on data mining and learning analytics in education.

According to Rani and Kautish (2018), a huge volume of data has been forming every second that necessitates development in repositories and database management. The authors argue that huge data are created in health care. In health care information comes from multiple sources. As a result, data volume is quite high but not all are equally relevant. It is necessary to understand which datasets are relevant and which are not. Unimportant data increases the risk of poor prediction. These problems are solved these days through machine learning and data mining techniques. The problem of medical management can be solved with the right use of data mining techniques. Data mining is also known as knowledge discovery. This paper reviews how the application of data mining techniques has evolved to its present form over the last decade. In

healthcare, as paper reviews, data mining helps to forecast the prognosis of chronic diseases like cancer, heart-related issues, and diabetes [16].

VII. CONCLUSION

Data mining is all about extracting useful rules and trending patterns from a large volume of datasets collected from various authentic sources. There are several data mining techniques that can be used to perform a job efficiently. It is true that there is no general technique for all types of database or for all sectors. Depending on the types of datasets appropriate techniques are used or to be used. Sometimes, hybrid techniques are used instead of single technique.

REFERENCES

1. Anderson (2019). Retrieved from: http://www.anderson.ucla.edu

2. Mohammed J. Zaki (2003). “DATA MINING TECHNIQUES”, August 2003

3. Sang Jun Lee, Keng Siau (2001). “A Review of Data Mining Techniques” Industrial Management & Data Systems 101/1, pp. 41-46

4. Dr. Rajni Jain, “Introduction to Data Mining Techniques”

5. Mostafa, Ashour. (2016). Review of Data Mining Concept and its Techniques. 10.13140/RG.2.1.3455.2729.

6. Ghuman, S., S. (2014). A Review of Data Mining Techniques.International Journal of Computer Science and Mobile Computing. IJCSMC, Vol. 3, Issue. 4, pg. 1401 – 1406

7. Gera, M. and Goel, S. (2015). Data Mining - Techniques, Methods and Algorithms: A Review on Tools and their Validity. International Journal of Computer Applications, Volume 113 – No. 18

8. Mirza, Mittal, and Zaman (2016). A Review of Data Mining Literature. International Journal of Computer Science and Information Security (IJCSIS),Vol. 14, No. 11

9. Manjarres, Sandova & Suárez (2018). Data mining techniques applied in educational environments: Literature review. Digital Education Review - Number 33, June 2018-http://greav.ub.edu/der/

10. Oluwaseun, Adelaja & Chaubey, Mani. (2019). DATA MINING CLASSIFICATION TECHNIQUES ON THE ANALYSIS OF STUDENT'S PERFORMANCE. 7. 17. 10.11216/gsj.2019.04.19671.

11. Rao, Ramana, and Ramkrishna (2019). Implementing the Data Mining Approaches to Classify the Images with Visual Words. International Journal of Recent Technology and Engineering (IJRTE), Volume-7 Issue-6S2.

12. Fathimath Zuha Maksood and Geetha Achuthan. Article: Analysis of Data Mining Techniques and its Applications. International Journal of Computer Applications 140(3): pp. 6-14.

13. Salem, Sayed-Mouchaweh, and Hassine (2017). A Review on Machine Learning and Data Mining Techniques for Residential Energy Smart Management

14. Tummala, Yashasree & Hemantha, Kumar & Kalluri, Hemantha Kumar (2018). A review on Data Mining & Big Data Analytics. International Journal of Engineering & Technology. 7. Pp. 92-94. 10.14419/ijet.v7i4.24.21863.

15. Sharma, Sharma, and Dwivedi (2017). Literature Review and Challenges of Data Mining Techniques for Social Network Analysis. Advances in Computational Sciences and Technology Volume 10, Number 5 (2017) pp. 1337-1354

16. Silva, Carla & Fonseca, Jose. (2017). Educational Data Mining: A Literature Review. 10.1007/978-3-319-46568-5_9.

Corresponding Author

Urvashi Sangwan*

Assistant Professor, Department of Computer Science and Engineering, Vaish College of Engg, Rohtak

usangwan@gmail.com