Getting to know Data Mining

Chandan.M.Bharadwaj
Department of Computer Science Engineering
Global Academy of Technology
Bengaluru, India

H.C.Prajwal
Department of Computer Science Engineering
Global Academy of Technology
Bengaluru, India

***Abstract*—Conventionally the term "mining" refers to the process of extraction of useful material from the surface of the earth, e.g. iron ore mining, gold mining etc. But concerning Computer Science it not only refers to the extraction of useful information from raw data or data warehouse but also the identification of patterns, finding anomalies and correlation with the large volume of data. This would help us to predict the outcome; hence making data mining also known as Knowledge Discovery or Knowledge Extraction. This paper surveys the basic concept of data mining, architecture, process flow, and emphasize data mining tools, application of data mining in various fields and challenges.**

Keywords—data mining; data warehouse; tools; application; challenges;

I. INTRODUCTION

Data Mining is a procedure of investigating and breaking down an enormous arrangement of data to find important examples and to create predictive models from them. It is an interdisciplinary field that includes ideas from different spaces, for example, Statistics, Machine Learning, Computing, Information Theory, Database Systems, and Pattern Recognition.

II. EVOLUTION OF THE TERM DATA MINING

The extraction of patterns from data has been in practice for centuries, earlier this was carried out manually by using Baye's theorem in the 18th century and regression analysis in the 19th century. With the advancement in the field of computer science, the data sets have also grown and complexity, the surge for advanced data handling techniques became a necessity. Hence there was the growth of Neural Networks, Cluster Analysis, Decision Trees, and Support Vector Machines. Data mining is the process of applying these methods to identify hidden patterns in enormous data sets and extract the useful and desired information from them.

III. NEED FOR DATA MINING

Data mining has vast applications in various domains as it can identify the insights and visions of data from data sets. Hence, it is an up-and-coming field for the present generation and has attracted enormous attention in the field of information industry and society. As today's world runs on data, vast amounts of data are available, and there is a greater need for converting raw data into valuable 2information. The extracted knowledge is utilized in various fields ranging from developing smart market decision, running accurate campaigns to analyzing customer behaviors and their insights.

IV. THE ARCHITECTURE OF DATA MINING SYSTEM

The data collected from various sources must be purified, merged, and selected before passing the data to the database or data warehouse server. As the cleaned data comes from varied sources and in different formats, the data can be incomplete and not accurate to be directly used for the data mining process. So, at first, data requires to be cleaned and then merged so that the additional information collected from various data sources is integrated well, with only data of interest passed to the server.

A. Database or Data Warehouse Server:

The database or data warehouse server holds the actual data ready to be processed. The data is retrieved as per the user's request.

B. Data Mining Engine

Data Mining Engine is the root and the major component of the data mining architecture. It contains several modules for the functionality of data mining tasks. It comprises of the accessories and software that assist in analyzing the processed data collected from various data sources and stored within the data warehouse.

C. Assessment of patterns in data

The Pattern assessment module in data mining is significantly answerable for looking through a pattern by utilizing threshold value. It works alongside the Data Mining Engine to concentrate on the recognizable proof and looking for a fascinating pattern.[2] This area uses different steady estimates that co-work with the data mining engine to scan for an energizing pattern from the data. It uses a threshold to filter out the discovered patterns from the available data. Pattern assessment module is coordinated alongside the Data Mining Engine relying upon the usage of the information mining strategy utilized. For effective outcomes during information mining, it is suggested to try pushing the evaluation of pattern stake into the mining procedure to confine the search to only fascinating patterns.

D. Graphical User Interface

The graphical user interface (GUI) module acts as a communication bridge between the data mining system and the user. The usage of this module is easy and efficient and is independent of the complexity of the process. This module helps the data mining framework when the client raises a question or a query and later presents the outcomes.

E. *Knowledge/Information Base*

The knowledge base is basic in the whole life pattern of information mining. It may be useful to direct the pursuit or assess the stake of the outcome designs. The information base may likewise contain client encounters and perspectives on the client steady during the time spent information mining. The information mining engine hence gets valuable contributions from the information, delivers a precise and solid outcome from the accessible information.

V. PROCESS OF DATA MINING

A. *Business understanding*

This phase includes the establishment of business and data mining goals.

- First, it is necessity to grasp business targets unmistakably and discover the needs of the business.
- Next, we need to survey the current circumstance by finding the assets, presumptions, imperatives, and other fundamental factors that necessary to be thought about.
- Then, from the business goals and current circumstances, we have to make data mining objectives to reach our business goals inside the current structure.
- Finally, a decent information mining plan has to be built up, to reach both business and data mining goals. The new setup should be as point by point as it is expected to be.

B. *Data understanding*

In this stage, we assess the data and check if it meets the set data mining objectives.

- The initial step manages to understand the information gathered from various accessible information sources, which help to acquaint with the data.
- This stage additionally incorporates some imperative exercises, for example, information stacking and information coordination which is basic to make the procedure of information assortment effective.
- In the subsequent stage, the "surface" or "gross" properties of obtained data should be examined and revealed.
- This further prompts the investigation of information by experiencing the data mining questions, which can be settled utilizing different exercises, for example, querying, reporting, and representation.

C. *Data Preparation*

The most important step of data mining is data preparation as it typically consumes the maximum amount of project time. Once we have the desired data to be analyzed, we carry out a few steps which include cleaning, construction of the dataset, converting it into an understandable form. Further, we can carry out the data exploration task to identify the patterns based on business understanding.

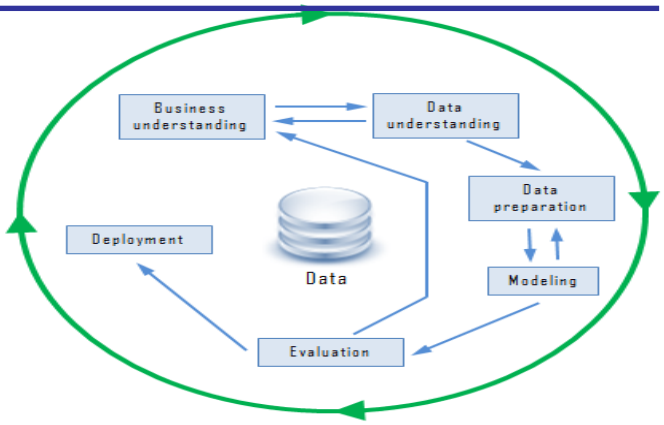

Fig. 1. Process of Data Mining

VI. DATA MINING TOOLS

The following are the popular data mining open source tools.

A. *RapidMiner*

This tool is written in Java programming language, and it offers an investigation of cutting-edge level through its format-based structure. Clients barely need to do any coding. RapidMiner is equipped for taking care of different undertakings like statistical modelling, predictive analytics, and visualization apart from data mining tasks. RapidMiner gives learning plans, models, and calculations from WEKA and R contents that make it all license, and it very well may be downloaded from Source Forge. It is one of the best business analytics software with all the data mining tasks bundled into one single suite.

B. *WEKA*

Weka was developed initially in a non-Java version for analyzing agricultural data. Later, the Java version was developed, and it became a powerful tool for different data mining applications like predictive modelling and data analysis. This software is free under the GNU General Public License, which is a significant advantage compared to RapidMiner. As it is open source under the GNU General Public License, it acts as a substantial advantage, compared to its counterpart like RapidMiner. The users can customize Weka to support most of the data mining jobs utilizing techniques like classification, clustering, regression, feature extraction, visualization, etc.

C. *R Programming*

Project R, which is a GNU project and available free of cost, is written in C, FORTRAN, and R Language. R language is used for writing lots of modules of the software itself and utilized for statistical computing and graphics. Data miners used R for developing statistical packages and analyzing the data. In recent years, the popularity of R has increased because of its ease of use and extensibility. R can be used for both linear and non-linear data modelling.

D. *Orange*

Orange, a Python-based, powerful open-source tool for data mining, is utilized by users for the purpose of knowledge extraction. It has robust visual programming and Python scripting framework attached to it. It also has application in

machine learning as well as bio informatics and text mining by adding add-ons. It is packed with features for data analytics.

E. NLTK

About language processing tasks, NLTK is one of the significant players. NLTK discovers applications in machine learning, data mining, sentiment analysis, and data scraping. Since it is written in Python, one can assemble applications on the top of it, by customizing it for small applications. NLTK finds handy application as a teaching tool, study tool, prototyping, and as a platform for high-quality research.

VII. DATA MINING APPLICATIONS

In this section, we have explored a few areas where data mining would be useful.

A. Financial Analysis:

It is observed that the finance and industry rely heavily on high-quality, reliable data for its transactions. In loan markets, financial and user data can be used for a variety of purposes, like predicting loan payments and determining credit ratings. Data mining methods make such tasks more manageable. Classification techniques help in segregation of crucial factors that influence customers' banking decisions from irrelevant ones. Further, it is also observed that multi-dimensional clustering techniques allow the identification of customers with similar loan payment behavior. Thus, data mining and analysis helps detect money laundering and other financial crimes.

B. Applications of data mining in the field of medicine

On account of the clinical examination, a patient's case is regularly investigated by tracking his clinical visits. This procedure assists with recognizing designs that have fruitful clinical treatments for different sorts of sicknesses endured by the patients. This data is then made accessible for specialists to assess the expense of treatment and improve the norm of administrations given by clinics. We additionally utilize measurements, information representation, and AI to decide and anticipate the volume of patients inside one class. The procedures are created to affirm that the patients get suitable consideration at whatever point required. It likewise helps human services and medication-based insurers to forestall misrepresentation cases.

C. Telecommunication Industry

Telecommunication is expanding and growing exponentially after the advent of the internet. Data mining can enable key industry players to improve their service quality to stay ahead in the game. Pattern analysis of spatiotemporal databases can play a huge role in mobile telecommunication, mobile computing, and web and information services. And techniques like outlier analysis can detect fraudulent users. Also, OLAP and visualization tools can help compare information, such as user group behavior, profit, data traffic, system overloads.

D. Intrusion Detection

Intrusion means any step taken that will put the confidentiality and integrity of a resource to risk. We can employ protective measures to prevent any intrusion, which includes effectively detecting any anomalies or deviation from normal behavior, adopting compulsory user authentication, avoid semantic errors and safeguarding information. It helps an analyst to distinguish activity from common everyday network activity. As we live in a technology-driven economy today, network administration is prone to security challenges. Network resources can face threats and actions that intrude on their confidentiality or integrity. Therefore, detection of intrusion has emerged as a crucial data mining practice.

E. E-commerce

One of the spaces in which data mining is broadly utilized in E-Commerce. Data mining is generally appropriate for this area as every single raw material required for data mining is promptly accessible: information records are accessible in mass, the electronic assortment gives dependable data, insight can easily be turned into action, and return on investment can be measured. The combination of online business and information mining altogether improves the outcomes and guide the user in creating information and settling on the right business choices. This combination successfully solves a few obstacles related with even data mining instruments, including the colossal exertion required in pre-preparing of the data before it very well may be used for mining and making the aftereffects of mining noteworthy.

F. Automated prediction of trends and behavior

With the exponential development of data mining, the way toward finding predictive information from massive data sets is automated. With the assistance of this, questions recent which required broad hands-on examination would now be addressed straightforwardly from the information rapidly. An excellent case of a predictive issue is focused on advertising. This procedure utilizes information on past promotional mailings to distinguish the objectives destined to expand quantifiable profit in future postings. Other predictive issues where data mining can be used are banking where they can anticipate bankruptcy where based on the previous data and other forms of default.

VIII. CHALLENGES FACED BY DATA MINING

In spite of the fact that data mining is viewed as incredible data assortment practice, it despite everything faces a few difficulties during its usage, such as difficulty in accessing data, the performance of the algorithms, several challenges related to mining methods, etc. These problems must be addressed by the companies for operational execution. Some of the challenges in data mining are as follows.

A. Difficulty in Accessing Data

Data miners, by and large, concurred that trouble in getting to data is because of the shortage of a thought/procedure for data. The significant inquiries are about how it tends to be frequently acquired, what sort of information is required, how quality will be guaranteed or improved, and the way it will be looked after and so forth. Once more, data miners recommend working straightforwardly with business clients to coordinate business issues with data prerequisites and to utilize this as the least difficult approach to begin building up a more extensive arrangement for data assortment and data availability.

B. Algorithm Performance

The most crucial aspect of data mining is algorithms. The performance of the data mining depends on the various factors such as algorithm used, and the mining method used. If the result of data mining is not satisfactory due to mining methods and algorithms, it will hamper the result and affect the end date resultantly.

C. Data Visualization

Data visualization might be a crucial procedure in data processing in the light of the fact that it is the most vital procedure that shows the yield in an adequate manner to the user. The data filtered should summarize the precise meaning of what it represents. But time and again, it is not easy to represent the data to the end-user in an error-free and straightforward manner. Since the input file and output information are relatively complex in nature, amazingly effective data visualization techniques must be applied for successfully fulfilling customer expectation.

D. Data Protection and Privacy

First and foremost, recurring issues for people and organizations are the security of data. The sphere of data mining typically poses critical data protection and data security issues.

E. Human interaction:

In the data mining process, user interaction plays a pivotal role. It describes how a user makes use of background knowledge in mining, visualization and finally comprehends the results of data mining. The interaction between the original data and the outcome of data mining result is vital in evaluating the efficiency of mining tasks.

F. Taking care of vulnerability, noise, or deficiency of data:

The error and noise in the data mining process may confuse the user, which may lead to the inception of incorrect patterns. Thus, data must undergo cleaning, processing, outlier detection, and removal of noise.

CONCLUSION

In this paper, we briefly discuss the concepts of data mining/knowledge mining and its tools, applications, and its limitation. Data mining or Knowledge Data Discovery is the computerized process of digging and examining huge arrangements of data and pulling out the meaning of the information from the data. It is applied viably within fields like business environment, weather forecasting, medication, transportation, healthcare, insurance, internal security, etc.

ACKNOWLEDGMENT

The authors of the paper would love to express their heartfelt thanks to our teachers Dr. N Guruprasad and Mr. Anand Panduranga, for their constant support and inspiration. They guided us in completing this paper.

REFERENCES

[1] Agarwal, S. (2013, December). Data mining: Data mining concepts and techniques. In 2013 International Conference on Machine Intelligence and Research Advancement (pp. 203-207). IEEE.

[2] Dunham, M. H. (2006). Data mining: Introductory and advanced topics. Pearson Education India

[3] Lakshmi, B. N., & Raghunandhan, G. H. (2011, February). A conceptual overview of data mining. In 2011 National Conference on Innovations in Emerging Technology (pp. 27-32). IEEE.

[4] Tan, P. N., Steinbach, M., & Kumar, V. (2016). Introduction to data mining. Pearson Education India

[5] Zhang, H. (2011, August). A short introduction to data mining and its applications. In 2011 International Conference on Management and Service Science (pp. 1-4). IEEE