International Journal of Engineering Sciences & Research Technology

(A Peer Reviewed Online Journal)
Impact Factor: 5.164

IJESRT

RESEARCHERID
THOMSON REUTERS

[Jogannagari *et al.*, 9(7): July, 2020]
IC™ Value: 3.00

ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

AN EMPIRICAL ANALYSIS ON DATA MINING SYSTEM AND ITS TECHNIQUES

Malla Reddy Jogannagari[*1], Benarji Tharini[2] & Haritha Chikoti[3]

[*1]Professor, Department of Computer Science & Engineering, Indur Institute of Engineering & Technology, Siddipet, Telangana State, India

[2]Assoc. Professor, Department of Computer Science & Engineering, Indur Institute of Engineering & Technology, Siddipet, Telangana State, India

[3]Project Trainee (R&D), Department of Computer Science & Engineering, Indur Institute of Engineering & Technology, Siddipet, Telangana State, India

DOI: https://doi.org/10.29121/ijesrt.v9.i7.2020.4

ABSTRACT

Data mining is an extraction of knowledge discovery from huge amount of data which is previously unknown and potentially useful for analytical processing and decision making. The other acronyms of data mining are such as Data archeology, Data dredging, Information harvesting and Business Intelligence. The various data mining techniques are used to find the hidden interestingness or new patter to store the data. These techniques and approaches of data mining can efficiently build the new environment for analyzing and predictions. This paper highlights data mining process and its various techniques to find the interestingness. Finally, concluded with its limitations. The objective of the paper is opens new horizons for researchers of forthcoming generations.

KEYWORDS: Data warehouse, Data Mining, Description Analysis, Prediction Analysis, Opportunities, Challenges.

1. INTRODUCTION

In the 1990, the organizations operations are dispersed into global arena. The organization executives are hopeless for gathering information to reaching the competitive edge and gain the bottom line. The operational computer systems which contains heterogeneous data at various locations. Due to the increased demand of the dynamic market, competitive pressure, culture and other similar factors motivated the organizations to review their storage structures, approaches and strategies. The organizations bring out the news ways for getting huge data for dynamic business.

The evaluation of data based systems started with file management in 1960. After 1970 onwards the concept of Database management system developed for data collection and processing with various features. General idea of data warehousing invented in 1980 to store the tremendous data with high velocity. The IBM researchers used buzzword “data warehousing” and given high popularity among the software industry. In 1990, the various data mining techniques used to extract the interestingness, the evolution of these techniques in the area of data warehousing with a long history.

Data Warehouse is a process of gathering and managing the data from heterogeneous sources to provide meaningful insights. It is core of Business Intelligence builds for data analysis and reporting. Data warehouse is a blend of technologies and provide the data for mining, which aids the strategic use of decision making. Data mining is the process of discovering anomalies patterns and correlations within the data set to predict outcomes. These technologies used in wide ranges of domains such as market analysis, financial, production control and fraud detection etc

Research paper focused on significance of data mining system and its techniques. The rest of paper organized is as follows. The Section 2 focused on the review literature of data mining. Section 3 describes the taxonomy of data mining and its spectrum. The section 4 focused on various techniques of data mining. Section 5 covered with the tools involved in Data mining. The research opportunities and its challenges discussed in Section 6. Finally, research paper concluded with Section 7.

ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7

2. REVIW OF LITERATURE

Over the few years no. of eminent researchers and theorists have worked on data mining systems. Their concepts and ideas can be useful the extend the knowledge in the area of data mining. The various researchers contributions is as follows.

Ansha. N. et al[2] identified the vital role and practicality of data mining in clinical research, instruction and human services of health care system. with a survey on Medical data.

Himani Rani. et al[4], published review on Prediction Analysis techniques of data mining which is combination of clustering and classification. In this review, various technique and trends of prediction are analyzed in a tabular form.

S. Nageshwari et al[5], implemented the data minig techniques on Student data sets. She compared the various classification techniques such as decision Tree, Neural Network, Support Vector Machine,. Naïve Bayes and K-Nearest Neighbor on quality perspective. Finally she concluded that, the Decision Tree and d neural Network provide the best accuracy.

Neelamadhab Padhy et al[6], focused the various techniques, approaches of data mining in different domains. Data warehouse have significant value to improve the effectiveness of managerial decisions making in the business environment.

Nelofar Rehmandiscussed that, the data mining is powerful technology with great potential to support the business operations and decision making in proactive and knowledge driven[7].

Nitesh Kumar Dokania et al [8], analyzed the various data mining techniques with comparative study. They proposed the "perfect" data mining model for dynamic data existed in the real time environment such as markets, financial, spatial and personal.

Rajkumar. S. et al [10], published research paper with aim of securing society from crimes. They identified there is need of advance system and new approaches for improving crime analytics for protecting society. In the paper data mining performs various criminal analysis and crime prediction towards the crime occurrences, hotspot & prediction, image patterns and mobile conversations etc. These are very much useful for the law enforcement officers to speed up the reduce the crime rate.

3. TAXONOMY OF DATA MINING AND ITS SPECTRUM

Data warehouse is a repository of tremendous data gathered from various sources and stored under unified schema located at single site. It is electronic storage of huge data designed for query and analysis. The huge amount of data in data warehouse gathered from different places of business operations such as marketing, sales, and finance, human resource etc. Data mining is extracts the interestingness from the data warehouse.

The Data warehouse collection of huge amount of data stored in high velocity and increased every day. Traditional database systems are inadequate to analyze this tremendous data. Data mining can analyze the data in different perspective and summarizing into meaningful information. It extracts and analyze the insights with help of query, reporting and decision making tools.

Recent data mining systems involved with data warehousing and advanced technologies such as Machine learning, artificial intelligence and statistical analysis. The business industry using broad range of data mining techniques to increase revenues, enhance the customer relationships, cost cutting and reduce the business risk to achieve the competitive advantage. The various advantages of the data warehousing is as follows.

Advantages:

***Enhance the Business Intelligence*:** Data warehouse consists with huge data gathered from multiple sources integrated as unified schema, the decision makers will no long need to rely on limited data. It provide the complete view of entire business such as market, inventory, financial, risk and sales.

ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7

Improved data quality and Consistency: A data warehouse converts the gathered data of multiple sources into unified format. Since the data from across the organization is standardized, every department will generate results in consistent format. It provides information on cross functional activities with more accurate and useful for strategic decisions.

Attain the high Returns on Investment [ROI] : The companies achieved higher revenues and cost savings with data warehouse than those doesn't invested in data warehouse.

Data Warehouse Saves Time: **Data warehouse is an integrated data repository with unified schema. Since critical data is available for decision making for all business users. It reduces the total around time for analysis and decision making.**

Better decision making: Data warehouse consists with complete data and provides better insights to users for effective decision making. Incomplete and inconsistent data leads to ambiguity decision making.

Historical Intelligence: A data warehouse consists of huge amount of historical data. It helps users to analyze different time periods and trends in order to make future predictions. Such data is not supported by transactional databases system.

Achieving the competitive advantage: Data warehouse enabled organizations achieve the competitive advantage than other organizations. These organizations make effective decision making and evaluate opportunities and risks rather than based on intuition.

Process of Knowledge discovery database constructed with sequence of the following steps and depicted with the figure 3.1.

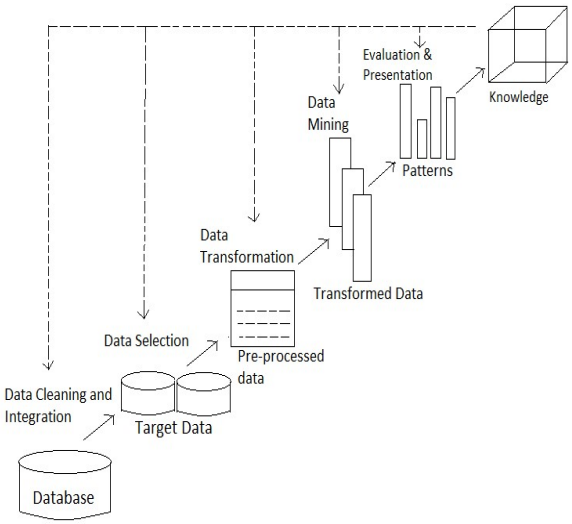

Figure3.1 Steps involved in Knowledge Discovery

The knowledge discovery is a high-level process of finding knowledge in data with the various steps such as data cleaning, integration, selection, transformation, loading and periodic refreshing.

Steps involved in the Data mining

Data Cleaning: Data cleaning is process of removing the noisy and irrelevant data from the large data set. In this process data redundancy, in consistency is removed and missing values are filled with appropriate values. All the data format are make into consistent and complete.

ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7

Data Integration: Heterogeneous data gathered from multiple sourced to be combined into common source. Data can be integrated with various migration tools, synchronization and ETL tools.

Data Selection*:* Data relevant to task analysis are selected from the databases. Select the enough quality of task oriented data to perform data selection.

Data transformation: This process can transform and consolidate the data into appropriate for mining process using data summary, aggregation and normalization procedures. In short, data is transformed into appropriate for data mining step.

Data mining*:* Based on the objective of data mining, appropriate task is selected such as classification, clustering, sequential pattern discover, association rule discovery and regression etc. These task can be chosen based on the need description and prediction.

Pattern evaluation: *Evaluation is process finding interestingness patterns and identified with given measures. This is a post processing step in KDD which interprets mined patter and relationships.*

Knowledge representation: This is final step of Knowledge Discovery in Databases (KDD). The mined knowledge is consolidated and represented to the user community in easy understandable format. In this phase various visualization techniques are applied on data mining results for better understanding.

Use of Discovered Knowledge: The extracted knowledge can improve the business operations with better decisions, reduce the costs, increase the profits with effective data mining techniques. Decisions are made based upon the availability of data as well as the environment situations such as certainty, uncertainty and risk.

Data mining system can be classified based on the kind of database such as relational, time series, stream, text, spatial, multimedia, and applications used, each require its own data mining technique etc.

4. VARIOUS TECHNIQUES OF DATA MINING

Data mining extract the interestingness from the huge amount of data. It retrieves patterns, associations, changes and anomalies from large data sets. It is a extraction of authentic, novel and potentially usable knowledge from existing data sets. Data mining tasks can be classified as the following diagram 4.1

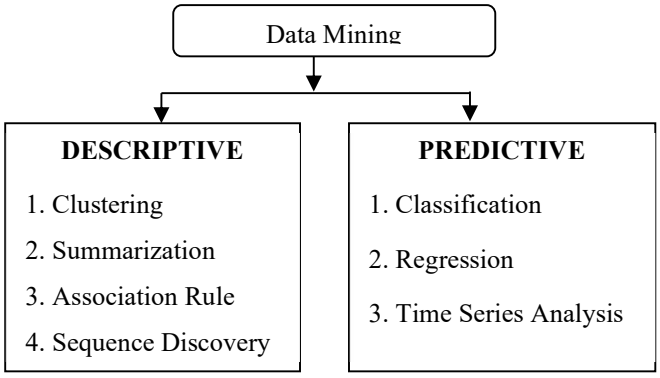

Figure 4.1 Classification of Data mining technique

Descriptive Analysis: Descriptive analysis represents the what happened past and present in understandable form[3]. It performs on general properties of the data stored in data base and classified into various other functionalities such as Clustering, Summarization, Association rule and Sequence discovery.

Clustering: Grouping the similar types object together is called as Clustering. In this technique the entire data partitioned into groups based on data similarity then assign the labels for each. The K-means clustering algorithm

ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7

is a hard splitting clustering widely used to its simplicity and speed. Another algorithm Y-mean clustering used for intrusion detection.

Summarization: Summarization is a key data mining technique used to find packed description of a dataset. The various statistical methods of mean, standard deviation are applied in the data analysis, visualization and automatic report generation.

Association Rule: Association analysis is the discovery interesting patterns, associations, correlations, or causal structures among set of items in relational databases. It is widely used in market basket or transaction data analysis.

***Sequence Discovery*:** Sequential pattern mining is a topic of data mining anxiety with finding marked apposite patterns between data examples where the morals are delivered in a sequence.

Predictive Analysis: Predictive analysis determines the future outcome rather than present[9]. It will focus on inferences on present data in order make predictions. It is classified into classification, Regression analysis and Time-series analysis based on perspective.

Classification*:* Classification data mining technique build the models from data with predefined classes. The derived model may represented in various forms such as IF.. THEN classification, decision trees and neural networks[1][9]. It can be used for predicting the class label of data objects. It is a machine learning method used to predict group bond for data patterns.

Regression: Regression is a statistical method used for numeric prediction that how many dependent variables effected with independent variable. It predicts profits, sales, temperatures and distances. For example the result of the student is depending on many parameters.

Time Series Analysis: Time series analysis is a trend analysis deals with times series of data over the time. The predication based on the data is in series at particular time intervals[8].

5. TOOLS INVOLVED IN DATA MINING

Rapid Miner: Rapid Miner is open source data science platform used for predictive analytics such as data mining, text analytics and machine learning. It integrates with various source data such as Microsoft Access, Excel, Teradata, Microsoft Sql and Oracle.

Weka: Weka environment provides Knowledge Analysis. The program written in Java

Oracle Data mining: It is Oracle product used by leading companies of analysis and predictions.

IBM SPSS Modeler: It provides modelers, text analytics and its state of the art visual interface prove to be extremely valuable.

Rattle: It is GUI open source package for data mining using R. analytical programming language.

KNIME: Konstanz Information Miner(KNIME) is a open source software that analyze the data model. It consists with integration, visualization and reporting features through its modular pipeline.

Apart from above other software's such as R Programming, Python, Orange, Teradata provides data mining features

ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7

6. RESEARCH OPPROTUNITIES AND ITS CHALLENGES

Data Mining is emerging research area used many techniques from other disciplines such as neural networks, fuzzy logic, mathematics, inductive logic programming. The other realms are machine learning, statistical analysis and data visualization are upcoming areas in the field of data analytics.

Research Opportunities:
Data mining is universal solution for different industries and sectors to meet the competitive edge[3]. Most of organizations and sectors such as transportation, healthcare, agriculture, education, social media and crime detection are using data mining applications for effective decision making.

Challenges:

- Most of the organizations maintaining data warehouse systems due to lack of awareness of data mining features. Without knowledge they hesitate the data mining
- Data warehouse is heterogeneous data set and affected noise, sometimes inconsistent. Efficient storage and processing is prerequisite for data warehouse.
- Gathering data from various sources leads to redundancy. Detecting and eliminating redundancy may improve the storage capacity.
- Today, Data warehouse features expanding in various domains. With lack of professional the organizations cant achieve the industry needs.
- The proper coordination is required in between the professional and management at different phases of analytics.
- Sometime management may not trust the outcomes/ insights of data mining. They don't know how data can generate such insights
- Security is most important for every technology. There is no exemption for data mining. Data warehousing heavily depending on cloud environment which consists with sensitive data of personal and business, It needs security from third party applications of unknown, can easily impose the risks into the enterprise networks with less security standards..

7. CONCLUSIONS

In today's advancement and ever changing economy, the information is a key resource for any organization to gain the market advantage. To gain this advantage the data mining has achieved amazing success in solving various business problems. This paper focused on the steps on knowledge discovery process and highlights the capabilities of data mining techniques. There is a need of extensive research in the area of data mining for changing business trends and opportunities

8. ACKNOWLEDGEMENTS

We sincerely acknowledge to all the faculty members of the Department of Computer Science & Engineering, Indur Institute of Engineering and Technology, Siddipet for their for their motivation and support in the period of Data Analysis.

REFERENCES

[1] Adelaja Oluwaseun Adebayo. et al, " Data Mining Classification Techniques on the Analysis of Student's Performance", Global Scientific Journals, Vol. 7(4), April, 2019
[2] Ansha. N et al, " A Survey on Medical Data by using Data Mining Techniques", International Journal of Science, Engineering and Technology(IJSETR)", Vol.7(1), Jan,2018.
[3] Deshpande.S.P. et al, " Data Mining System and Applications : A Review", International Journal of Distributed & Parallel System (IJDPS), Vol. 1(1), Sep, 2010.
[4] Himani Rani et al, "Prediction Analysis Technique of Data Mining : A Review", International Journal of Computer Science and Mobile Computer", Vol. 8(5), May,2019.
[5] Nageshwari. S et al, "Comparison of Classification Techniques on Data Mining", International Journal of Emerging Technology and Innovative Engineering", Vol. 5(5), May, 2019.
[6] Neelamadhab Padhy, " The Survey of Data Mining Applications", International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol. 2(3), June, 2012.

ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7

[7] Nelofar Rehman, " Data Mining Techniques Methods, Algorithms and Tools", International Journal of Computer Science & Mobile Computing, Vol. 6(7),July, 2017
[8] Nilesh Kumar Dokania et al," Comparative study of various Techniques in Data Mining", International Journal of Engineering Sciences & Research", Vol. 7(5), May.2018.
[9] Poonam Chaudhary," Data Mining System, Functionalities and Applications : A Radical Review", International Journal of Innovations in Engineering and Technology (IJIET), Vol 5(2), April, 2015.
[10] Rajkumar.S. et al, "Crime Analysis and Prediction using Data Mining Techniques", International Journal of Recent Trends in Engineering & Research", Special Issue, Mar, 2019.