International Journal of Trend in Scientific Research and Development (IJTSRD)
Volume 4 Issue 4, June 2020 Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

Overview of Data Mining

Rupashi Koul

Department of Computer Science Engineering, Dronacharya College of Engineering, Gurugram, Haryana, India

ABSTRACT

Data mining is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems.[1] Data mining is an interdisciplinary sub field of computer science and statistics with an overall goal to extract from a data set and transform the information into a comprehensible structure for further use.[1][2][3][4] The process of digging through data to discover hidden connections and predict future trends has a long history. Sometimes referred to as 'knowledge discovery' in databases, the term data mining wasn't coined until the 1990s. What was old is new again, as data mining technology keeps evolving to keep pace with the limitless potential of big data and affordable computing power. Over the last decade, advances in processing power and speed have enabled us to move beyond manual, tedious and time-consuming practices to quick, easy and automated data analysis. The more complex the data sets collected, the more potential there is to uncover relevant insights.

KEYWORDS: *database, data mining, techniques*

How to cite this paper: Rupashi Koul "Overview of Data Mining" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-4 | Issue-4, June 2020, pp.1333-1336, URL: www.ijtsrd.com/papers/ijtsrd31368.pdf

IJTSRD31368

I. INTRODUCTION

The manual extraction of patterns from data has occurred for centuries. Early methods of identifying patterns in data include Bayes' theorem (1700s) and regression analysis (1800s). The proliferation, ubiquity and increasing power of computer technology have dramatically increased data collection, storage, and manipulation ability. As data sets have grown in size and complexity, direct data analysis has increasingly been augmented with indirect, automated data processing, aided by other discoveries in computer science, specially in the field of machine learning, such as neural networks, cluster analysis, genetic algorithms (1950s), decision trees and decision rules (1960s), and support vector machines (1990s). Data mining is the process of applying these methods with the intention of uncovering hidden patterns[5] in large data sets. It bridges the gap from applied statistics and artificial intelligence to database management by exploiting the way data is stored and indexed in databases to execute the actual learning and discovery algorithms more efficiently, allowing such methods to be applied to ever-larger data sets. mining is widely used in diverse areas. There are a number of commercial data mining system available today and yet there are many challenges in this field. Some of the areas in which data mining is used is as follows:

A. Retail Industry

Data Mining has its great application in Retail Industry because it collects large amount of data from on sales, customer purchasing history, goods transportation, consumption and services. It is natural that the quantity of data collected will continue to expand rapidly because of the increasing ease, availability and popularity of the web. Data mining in retail industry helps in identifying customer buying patterns and trends that lead to improved quality of customer service and good customer retention and satisfaction.

B. Telecommunication Industry

Today the telecommunication industry is one of the most emerging industries providing various services such as fax, pager, cellular phone, internet messenger, images, e-mail, web data transmission, etc. Due to the development of new computer and communication technologies, the telecommunication industry is rapidly expanding. This is the reason why data mining is become very important to help and understand the business. Data mining in telecommunication industry helps in identifying the telecommunication patterns, catch fraudulent activities, make better use of resource, and improve quality of service.

C. Education

There is a new emerging field, called Educational Data Mining, concerns with developing methods that discover knowledge from data originating from educational Environments. The goals of EDM are identified as predicting students' future learning , studying the effects of educational support, and advancing scientific knowledge about learning. Data mining can be used by an institution to take accurate decisions and also to predict the results of the student. With the results the institution can focus on what to teach and how to teach. Learning of the students can be captured and used to develop techniques to teach them.

D. CRM

Customer Relationship Management is all about acquiring and retaining customers, also improving customers' loyalty

and implementing customer focused strategies. To maintain a proper relationship with a customer a business need to collect data and the information. This is where data mining plays its part. With data mining technologies the collected data can be used for analysis. Instead of being confused where to focus to retain customer, the seekers for the solution get filtered results.

E. Fraud Detection

Billions of dollars have been lost to the action of frauds. Traditional methods of fraud detection are and complex. Data mining aids in providing meaningful patterns and turning data into information. Any information that is valid and useful is knowledge. A perfect fraud detection system should protect information of all the users. A supervised method includes collection of sample records. These records are classified fraudulent or non-fraudulent. A model is built using this data and the algorithm is made to identify whether the record is fraudulent or not.

F. Intrusion Detection

Any action that will compromise the integrity and confidentiality of a resource is an intrusion. The defensive measures to avoid an intrusion includes user authentication, avoid programming errors, and information protection. Data mining can help improve intrusion detection by adding a level of focus to anomaly detection. It helps an analyst to distinguish an activity from common everyday network activity. Data mining also helps extract data which is more relevant to the problem.

II. PROCESS OF DATA MINING

The data mining process is divided into two parts i.e. Data and Data Mining. Data involves data cleaning, data integration, data reduction, and data transformation. The data mining part performs data mining, pattern evaluation and knowledge representation of data.

A. Data Cleaning

Data cleaning is the first step in data mining. It holds importance as dirty data if used directly in mining can cause confusion in procedures and produce inaccurate results. Basically, this step involves the removal of noisy or incomplete data from the collection. Many methods that generally clean data by itself are they are not robust.

B. Data Integration

When multiple heterogeneous data sources such as databases, data cubes or files are combined for analysis, this process is called data integration. This can help in improving the accuracy and speed of the data mining process. Different databases have different naming conventions of variables, by causing redundancies in the databases. Additional Data Cleaning can be performed to remove the redundancies and inconsistencies from the data integration without affecting the reliability of data.

C. Data Reduction

This technique is applied to obtain relevant data for analysis from the collection of data. The size of the representation is much smaller in volume while maintaining integrity. Data Reduction is performed using methods such as Naive Bayes, Decision Trees, Neural network, etc.

D. Data Transformation

In this process, data is transformed into a form suitable for the data mining process. Data is consolidated so that the mining process is more efficient and the patterns are easier to understand. Data Transformation involves Data Mapping and code generation process.

E. Data Mining

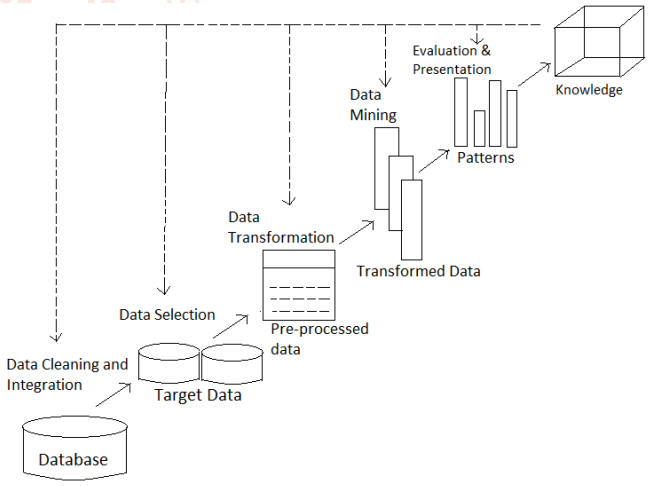
Data Mining is a process to identify interesting patterns and knowledge from a large amount of data. In these steps, intelligent patterns are applied to extract the data patterns. The data is represented in the form of patterns and models are structured using classification and clustering techniques.

F. Pattern Evaluation

This step involves identifying interesting patterns representing the knowledge based on measures. Data and visualization methods are used to make the data understandable by the user.

G. Knowledge Representation

Knowledge representation is a step where data visualization and knowledge representation tools are used to represent the mined data. Data is visualized in the form of reports, tables, etc.



III. TYPES OF DATA MINED

A. Flat files:

Flat files is defined as data files in text form or binary form with a structure that can be easily extracted by data mining algorithms. Data stored in flat files have no relationship or path among themselves, like if a relational database is stored on flat file, then there will be no relations between the tables. Flat files are represented by data dictionary.

B. Relational Database:

A Relational database is defined as the collection of data organized in tables with rows and columns. Physical schema in Relational databases is a schema which defines the structure of tables. Logical schema in Relational databases is a schema which defines the relationship among tables.

C. Data Warehouses:

A is defined as the collection of data integrated from multiple sources that will and There are three types of: Enterprise, Data Mart and Virtual Warehouse. Two approaches can be used to update data in Data Warehouse: Query-driven Approach and Update-driven Approach.

D. Databases:
databases is a collection of data organized by time stamps, date, etc to represent transaction in databases. This type of database has the capability to roll back or undo its operation when a transaction is not completed or committed. It is highly flexible system where users can modify information without changing any sensitive information.

E. Multimedia databases:
Multimedia databases consists audio, video, images and text media. They can be stored on Object-Oriented Databases. They are complex information in a formats.

F. Spatial Databases:
Spatial databases store geographical information. can store data in the form of coordinates, topology, lines, polygons, etc.

G. Time Series Databases:
Time series databases contains stock exchange data and user logged activities. handle array of numbers indexed by time, date, etc. It requires real-time analysis.

H. WWW:
WWW refers to World wide web which is a collection of documents and resources like audio, video, text, etc which are identified by Uniform Resource Locators (URLs) through web browsers, linked by HTML pages, and accessible via the Internet network. It is the most heterogeneous repository as it collects data from multiple resources. It is dynamic in nature as volume of data is continuously increasing and changing.

IV. DATA MINING TECHNIQUES

Data mining is highly effective and some techniques used for data mining are as follows:

A. CLASSIFICATION ANALYSIS
This analysis is used to retrieve important and relevant information about data, and metadata. It is used to classify different data in different classes. Classification is similar to clustering in a way that it also segments data records into different segments called classes. But unlike clustering, here the data analysts would have the knowledge of different classes or cluster. So, in classification analysis you would apply algorithms to decide how new data should be classified.

B. ASSOCIATION RULE LEARNING
It refers to the method that can help you identify some interesting relations (dependency modeling) between different variables in large databases. This technique can help you unpack some hidden patterns in the data that can be used to identify variables within the data and the concurrence of different variables that appear very frequently in the . rules are useful for examining and forecasting customer behavior. It is highly recommended in the retail industry analysis. This technique is used to determine shopping basket data analysis, product clustering, catalog design and store layout. In IT, programmers use association rules to build programs capable of machine learning.

C. ANOMALY OR OUTLIER DETECTION
This refers to the observation for data items in a that do not match an expected pattern or an expected behavior. Anomalies are also known as outliers, novelties, noise, deviations and exceptions. Often they provide critical and actionable information. An anomaly is an item that deviates considerably from the common average within a or a combination of data. These types of items are statistically aloof as compared to the rest of the data and hence, it indicates that something out of the ordinary has happened and requires additional attention. technique can be used in a variety of domains, such as intrusion detection, system health monitoring, fraud detection, fault detection, event detection in sensor networks, and detecting disturbances. Analysts often remove the anomalous data from the top discover results with an increased accuracy.

D. CLUSTERING ANALYSIS
The cluster is actually a collection of data objects; those objects are similar within the same cluster. That means the objects are similar to one another within the same they are rather they are dissimilar or unrelated to the objects in other groups or in other clusters. Clustering analysis is the process of discovering groups and clusters in the data in such a way that the degree of association between two objects is highest if they belong to the same group and lowest otherwise. result of this analysis can be used to create customer profiling.

E. REGRESSION ANALYSIS
In statistical terms, a regression analysis is the process of identifying and analyzing the relationship among variables. It can help you understand the characteristic value of the dependent variable changes, if any one of the independent variables is varied. This means one variable is dependent on another, but it is not vice versa. is generally used for prediction and forecasting.

V. BENEFITS AND DISADVANTAGES OF DATA MINING

There are several types of benefits and advantages of data mining systems. Some of them are as follows:

- One of the common benefits that can be derived with these data mining systems is that they can be helpful while predicting future trends. And that is quite possible with the help of technology and behavioral changes adopted by the people.
- Data mining helps organizations to make the profitable adjustments in operation and production.
- The data mining is a cost-effective and efficient solution compared to other statistical data applications.
- Most parts of the data mining process is basically from information gathered with the help of marketing analysis. With the help of such marketing analysis, one can also find out those fraudulent acts and products available in the market. Moreover, with the help of it one can understand the importance of accurate information.
- It can be implemented in new systems as well as existing platforms. is the speedy process which makes it easy for the users to analyze huge amount of data in less time.

Data mining technology is something that helps one person in their and that is a process wherein which all the factors of mining is involved precisely and while the involvement of these mining systems, one can come across several disadvantages of data they are as follows:

- There are chances of companies may sell useful information of their customers to other companies for money.

- Many data mining analytics software is difficult to operate and requires advance training to work on.
- Different data mining tools work in different manners due to different algorithms employed in their design. Therefore, the selection of correct data mining tool is a very difficult task.
- The data mining techniques are not accurate, and so it can cause serious consequences in certain conditions.

VI. CONCLUSION

Data Mining is an iterative process where the mining process can be refined, and new data can be integrated to get more efficient results. Data Mining meets the requirement of effective, and flexible data analysis. It can be considered as a natural evaluation of information technology. As a knowledge discovery process, data preparation and data mining tasks complete the data mining process. Data mining processes can be performed on any kind of data discussed in the above section. Finally, the bottom line is that all the techniques help in the discovery of new creative things. At the end of this paper about data mining, one can clearly understand the areas of applications, types of source data, process, techniques, and benefits with its own limitations. Therefore, after reading all the above-mentioned information about data mining one can determine its credibility and feasibility even better.

References

[1] "Data Mining Curriculum". ACM SIGKDD. 2006-04-30. Retrieved 2014-01-27.

[2] ^ Clifton, Christopher (2010). "Encyclopædia Britannica: Definition of Data Mining". Retrieved 2010-12-09

[3] ^ Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome (2009). "The Elements of Statistical Learning: Data Mining, Inference, and Prediction". Archived from the original on 2009-11-10. Retrieved 2012-08-07

[4] ^ Han, Kamber, Pei, Jaiwei, Micheline, Jian (June 9, 2011). Data Mining: Concepts and Techniques (3rd ed.). Morgan Kaufmann. ISBN 978-0-12-381479-1.

[5] Kantardzic, Mehmed (2003). Data Mining: Concepts, Models, Methods, and Algorithms. John Wiley & Sons. ISBN 978-0-471-22852-3. OCLC 50055336.