International Journal of Research
Available at https://edupediapublications.org/journals
e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 05 Issue 12
April 2018

Assessment on Progression, Techniques of Data Mining and Its Applications

Morigadi vishwashanthi & Botla Mamatha
Assistant professor, Dept .of CSE Geethanjali College of Engineering and Technology

vishwashanthi5b9@gmail.com
Assistant professor, Dept .of CSE Geethanjali College of Engineering and Technology

botlamamatha.74@gmail.com

Abstract: *Data mining is the process of analyzing data from different views and summarizing it into useful data. "Data mining, also popularly referred to as knowledge discovery from data (KDD), is the automated or convenient extraction of patterns representing knowledge implicitly stored or captured in large databases, data warehouses, the Web, other massive information repositories or data streams.". This paper provides a survey on various data mining techniques such as classification, clustering, regression, summarization and so on. This paper also discusses some of the data mining applications, additionally gift data processing primitives, from that data processing question languages will be designed. Problems concerning a way to integrate an information mining system with a database or data warehouse are mentioned. Additionally to finding out a classification of information mining systems, and its difficult analysis problems for building data processing tools of the long run.*

Keywords: **knowledge discovery in data, data mining application, descriptive model, predictive model**.

I. INTRODUCTION

Data mining, discovering of hidden predictive information from large data sets and it is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. Web could be a large repository of data that grows at a quick pace. The extreme growth of data evolves several new challenges for web researchers that embody among alternative things, high knowledge spatial property and extremely volatile and constantly evolving content. Be grateful to this, it's become more and more necessary to form new and improved approaches to ancient data processing techniques may be applied for the net mining. Automatically extracting helpful data could be key difficult problems in web data processing. The billions of sites created are generated dynamically by underlying web information service engines mistreatment HTML or XML. However, searching, comprehending, and mistreatment the semi structured data keep on the online poses a major challenge as a result of this knowledge is additional refined and dynamic than the data that business info systems store. The mining knowledge varies from structured to unstructured. Data processing chiefly deals with structured knowledge organized in an exceedingly info whereas text mining chiefly handles unstructured knowledge. Web mining lies in between and copes with semi structured knowledge and/or unstructured knowledge. Web mining entails artistic use of knowledge mining and/or text mining techniques and its distinctive approaches. Mining the net knowledge is one among the foremost difficult tasks for the information mining and data management students as a result of there are vast heterogeneous, less structured knowledge accessible on the online and that we will simply weak with knowledge. Because the web reaches its full potential, however, we have a tendency to should improve its services, build it additional approachable, and increase its usability. As researchers still develop data processing techniques, we have a tendency to believe this technology can play a progressively important role in meeting the challenges of developing the intelligent web.

II. DATA MINING PROCESS

Data mining is also known as Knowledge Discovery in Database, refers to finding or "mining" knowledge from large amounts of data. Data mining techniques are used to operate on large volumes of data to discover hidden patterns and relationships helpful in decision making. So, many people use the term "knowledge discovery in data" or KDD for data mining [1].

In Data mining, Knowledge extraction or discovery is done in seven sequential steps as in Fig 1.

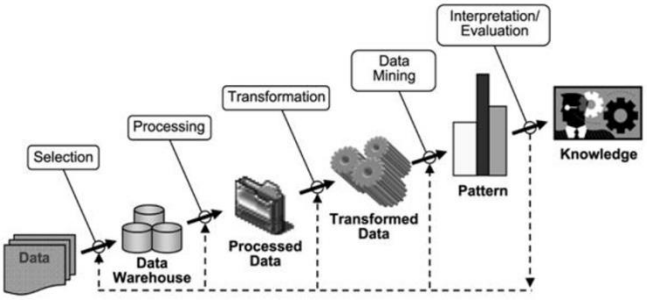

Figure 1 Data Mining Process

i) Data cleaning: This is the first step to eliminate noise data and irrelevant data from collected raw data.

ii) Data integration: At this step, various data sources are combined into meaningful and useful data.

iii) Data Selection: Here, data relevant to the analysis are retrieved from various resources.

iv) Data transformation: In this step, data is converted or consolidated into required forms for mining by performing different operations such as smoothing, normalization or aggregation.

v) Data Mining: At this step, various clever techniques and tools are applied in order to extract data pattern or rules.

vi) Pattern evaluation: At this step, Attractive patterns representing knowledge are identified based on given measures.

vii) Knowledge representation: This is the last stage in which, visualization and knowledge representation techniques are used to help users to understand and interpret the data mining knowledge or result.

The goal of knowledge discovery and data mining process is to discover the patterns that are unknown among the huge set of data and interpret useful knowledge and information.

III. DATA MINING TECHNIQUES

Data mining process is extraction of information from large data sets and transforms it into some understandable form for further uses. So it helps to achieve the specific objectives. The goal of a data mining effort is normally either to create a descriptive model or a predictive model[7]. A Descriptive model presents the data in concise form which is essentially a summary of the data points, finds patterns in the data and understands the relationships between attributes represented by the data. The Descriptive model includes tasks such as Clustering, Association Rules, Summarizations, and Sequence Discovery. The predictive model works by making a prediction about values of data, which uses known results found from different datasets [5]. The Predictive data mining model includes classification, prediction, regression and analysis of time series as in figure 2

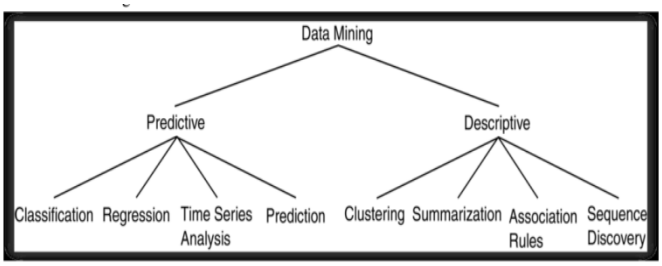

Figure 2 Data Mining Techniques

- **Classification:** Classification is the most commonly applied data mining technique, which employs a set of pre-classified examples to develop a model that can classify the population of records at large [3]. This approach frequently employs decision tree or neural network-based classification algorithms. The common characteristics of classification tasks are as supervised learning, categories dependent variable and assigning new data to one of a set of well-defined classes. Classification technique is used in customer segmentation, modeling businesses, credit analysis, and many other applications. E.g., classify countries based on population, or classify bikes based on mileage.
- **Regression:** Regression is another Predictive data-mining model is also known as supervised learning technique. This technique analyzes the dependency of some attribute values, which is dependent upon the values of other attributes mainly, present in same item. In the regression techniques target value are known. For example, you can predict the child's behavior based on family history.
- **Time Series data analysis:** Time-series database uses sequences of values or events obtained over repeated measurements of time. The values are typically measured at equal time interval such as hourly, daily, weekly. A sequence database is any database that consists sequence of ordered events, sometimes having concrete notions of time[6]. For example, Web page traversal sequences and customer shopping transaction sequences are sequence data, but they may not be time-series data.
- **Prediction:** This technique discovers the relationship between independent variables and the relationship between dependent and independent variables. The prediction is to predict a future state, rather than a current one [6]. Its applications include obtaining forewarning of natural disasters (flooding, hurricane, snowstorm, etc), epidemics, stock crashes, etc. As another example, the sales volume of computers accessories can be forecasted based on the number of computers sold in the past few months.
- **Clustering:** Clustering is a collection of similar data objects. Dissimilar object is another cluster. It is way finding similarities between data according to their characteristic Clustering can be considered as identification of similar classes of objects. By using clustering techniques we can further identify dense and sparse regions in object space and can discover overall distribution pattern and correlations among data attributes. Classification approach can also be used for effective means of distinguishing groups or classes of object but this method is expensive so clustering can be used as preprocessing approach for attribute subset selection and classification.

For example, image processing, pattern recognition, city planning. Astronomy - aggregation of stars, galaxies, or super galaxies

- **Summarization:** Summarization is referred as the abstraction or generalization of data. The summarization technique maps data into subsets with simple descriptions. The summarized data set gives general overview of the data with aggregated information. Simple summarization methods such as tabulating the mean and standard deviations are often applied for data analysis, data visualization and automated report generation. For example: length can be summarized as meters, centimeters or millimeters.
- **Association:** The Association technique is used to extract the relationships between attributes and items. In

this technique, the presence of one model implies the presence of another model i.e. item is related to another in terms of cause-and-effect. This is common in establishing a form of statistical relationships among different interdependent variables of data mining; association rules are useful for analyzing and predicting customer behavior. They also play an important role in shopping basket data analysis, product clustering, catalog design and store layout. The association rules are also build by programmers can be used to build programs capable of machine learning.

- **Sequence Discovery:** Uncovers correlation among data. It is set of object each associated with its own timeline of events. For example, scientific experiment, natural disaster and analysis of DNA sequence.

IV. DATA MINING APPLICATIONS

The Data mining applications are widely used in diverse areas such as retail stores, hospitals, banks, and insurance companies [2]. Many domains like health care, finance insurance, retail stores combines the data mining applications with statistics, pattern recognition, and other important tools to perform data analytics. Data mining is used primarily for decision making.

i. Medicare and health care: Data mining in medicine enables to characterize patient activities to see incoming office visits.Data mining helps identify the patterns of successful medical therapies for different illnesses.

ii. Education: Educational Data Mining is a blooming field which provides knowledge from educational Environment data. The goals of EDM are identified as predicting students' learning behavior, emotions and skills [3]. This study improves the educating methods by understanding the ward and to take accurate decisions respectively

iii. Market Basket Analysis: Market basket analysis is a technique that uses association rule mining to understand the purchasing behavior of the customer. It also allows the seller to understand his business, customer's needs and to make profitable change accordingly.

iv. Financial Banking: Data mining can contribute to solving business problems in banking and finance by finding patterns, causalities, and correlations in business information and market prices. The managers may find this information for better segmenting, targeting, acquiring, retaining and maintaining a profitable customer.

v. Research Analysis: Data mining is very useful in data pre-processing and integration of databases. Data mining allows the researchers to identify co-occurring sequences and the correlation between any activities .Data visualization and visual data mining help the researcher with a clear view of the data.

vi. Fraud Detection: The traditional fraud detection methods are expensive, time consuming and complex. Data mining aids in providing meaningful patterns and turning data into information.Vaild and useful information is called as knowledge. The results are categorized into fraudulent or non-fraudulent.

vii. Transportation: Data mining helps determine the distribution schedules among warehouses and outlets and analyze loading patterns.

viii. Agriculture: Data mining is emerging technology in agriculture field for crop yield analysis with respect to four parameters namely year, rainfall, production and area of sowing. Yield prediction is a very important agricultural problem that remains to be solved based on the available data. The yield prediction problem can be solved by employing Data Mining techniques such as K Means, K nearest neighbor (KNN), Artificial Neural Network and support vector machine

ix. Cloud Computing: Data Mining techniques are used in cloud computing. The implementation of data mining techniques through Cloud computing will allow the users to retrieve meaningful information from virtually integrated data warehouse that reduces the costs of infrastructure and storage. Cloud computing uses the Internet services that rely on clouds of servers to handle tasks. The data mining technique in Cloud Computing helps to perform efficient, reliable and secure services for their users.

V. CONCLUSION

According to the techniques of data mining listed above, it is learned that this a powerful and essential technique for performing manipulation of data that is data mining gives proper and targeted outcome from large and vastly growing data worldwide. This paper discusses the idea of data mining, the process of KDD, different techniques such as clustering, association, classification, prediction and so on. We also discussed some insights of the data mining applications.

REFERENCES

[1]. Aarti Sharma et al, "Application of Data Mining – A Survey Paper", International Journal of Computer Science and Information technologies', Vol. 5 (2), 2014.
[2]. Smita, Priti and Sharma, "Use of Data Mining in Various Field: A Survey Paper" IOSR Journal of Computer Engineering, 8727Volume 16, Issue 3, Ver. V (May-Jun. 2014)
[3]. Brijesh Kumar Baradwaj, Saurabh Pal" Mining Educational Data to Analyze Students Performance" (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 2, No. 6, 2011
[4]. J. Han and M. Kamber. "Data Mining, Concepts and Techniques", Morgan Kaufmann, 2000.
[5]. Nikita Jain, Vishal Srivastava "DATA MINING TECHNIQUES: A SURVEY PAPER" IJRET: International Journal of Research in Engineering and Technology, Volume: 02 Issue: 11 | Nov-2013.
[6]. Prof. Dr. Wolfgang Karl Hardle," Time Series Data Mining Methods: A Review", Berlin, March 25, 2015.
[7]. Pradnya P. Sondwale, "Overview of Predictive and Descriptive Data Mining Techniques" IJARCSSE, Volume 5, Issue 4, April 2015
[8]. Data Mining: Concepts and Techniques, Third Edition (The Morgan Kaufmann Series in Data Management System`s) 3rd Edition

[9]. R. Chau, C. Yeh and K. Smith, Personalized multilingual web content mining, KES (2004)
[10]. B. Liu and K. Chang, Editorial: Special issue on web content mining, SIGKDD Explorations 6(2) (2004)
[11]. Ricardo Baeza-Yates and Alessandro Tiberi. ―Extracting semantic relations from query logsproceeding for ACM SIGKDD international conference on Knowledge discovery and data mining, 2007.
[12]. P. Kolari and A. Joshi, Web mining: Research and practice, Comput. Sci. Eng.July/August (2004)
[13]. Ramakrishna, Gowdar et al Web Mining: Key Accomplishments, Applications and Future Directions, in the International Conference on Data Storage and Data Engineering 2010.

About the authors:

Morigadi vishwashanthi working as an Assistant professor in Department of CSE at Geethanjali college of Engineering and Technology affiliated to JNTU Hyderabad. She has 3 years teaching experience. Her Research interest includes Data Mining and Cloud Computing.

Botla Mamatha working as an Assistant professor in Department of CSE at Geethanjali college of Engineering and Technology affiliated to JNTU Hyderabad. She has 3 years teaching experience. Her Research interest includes Data Mining and Cloud Computing.