Knowledge Discovery from Data Mining Techniques

Nitin Kumar[1]
[1]Assistant Professor,
Department of Computer Science,
Shobhit University Gangoh (U.P)

Sumika Jain[2]
[2] Assistant Professor,
Department of Computer Science,
Shobhit University Gangoh (U.P)

Kuldeep Chauhan[3]
[3] Assistant Professor,
Department of Computer Science,
Shobhit University Gangoh (U.P)

***Abstract*—An evolving topic in today's era is Data Mining and Knowledge Discovery. Data mining and knowledge discovery in databases is attracting a lot of researchers, industry persons, academicians. Why this area is so emerging? This article provides an overview of this emerging field, gives an overview that how data mining and knowledge discovery in databases are related to each other and also to other related fields, such as machine learning, statistics, and databases. The article also mentions particular real-world applications, specific data-mining techniques, challenges involved in real-world applications of knowledge discovery, and current and future research directions in the field**

Keywords: Data Mining, Knowledge Discovery, Data Cleaning, Data Warehousing.

1. INTRODUCTION

Data is raw material of information that can be understood as any facts, numbers, or text which can be processed by machines. Information is the data that has been given some meaning nu way of relational connections. For ex data collected from sales transaction can be used to analyze sales trends of particular years. Knowledge is application of data and information.it can be considered as general awareness of information, facts, ideas, truth or principle.

Across a wide variety of fields, data are being collected and accumulated at a dramatic pace. There is an urgent need for a new generation of computational theories and tools to assist humans in extracting useful information (knowledge) from the rapidly growing volumes of digital data. These theories and tools are the subject of the emerging field of knowledge discovery in databases.

At an abstract level, the KDD field is concerned with the development of methods and techniques for making sense of data. The basic problem addressed by the KDD process is one of mapping low-level data (which are typically too voluminous to understand and digest easily) into other forms that might be more compact (for example, a short report), more abstract (for example, a descriptive approximation or model of the process that generated the data), or more useful (for example, a predictive model for estimating the value of future cases). At the core of the process is the application of specific data-mining methods for pattern discovery and extraction.

Data Mining is basically used today by most of the companies with a very strong consumer focus — retail, financial, communication, and marketing organizations, to "drill down" into their transactional data and determine pricing, client preferences and product related information, impact on the sales, client satisfaction and corporate profits. With the help of data mining, a retailer can use point-of-sale records of client purchases to develop products and promotions to appeal to specific client segments.

Knowledge discovery and data mining have become areas of growing significance because of the recent increasing demand for KDD techniques, including those used in knowledge acquisition, machine learning, databases, statistics, data visualization, and high performance computing. Knowledge discovery and data mining can be very useful for the field of Artificial Intelligence in many areas, for example education, industry, commerce, government, and so on. The relation between Knowledge and Data Mining, and Knowledge Discovery in Database (KDD) process are presented in the paper. Data mining theory, Data mining tasks, Data Mining technology and Data Mining challenges are also proposed.

The rules of Data mining are around a lot of functional elements. These functional elements also include the following:

Statistics: This discipline is allocated completely to the analysis of data. Many mathematical models are framed and the data is used as input for pattern analysis. This is used for association rules verification in data mining process.

Machine learning: In this area, the data sets are analyzed for models with statistical inferences and computational parameters. Most of the mining algorithms have machine learning ground work in them. Database technology: In this phenomenon, the prescribed data set is optimized using different techniques like compression, query compounding and data set indexing and are mined for relevant unknown patterns.

Information theory: This discipline is applied in the sector of communication where the information that is synthesized and processed are quantitatively measured by employing a technique where the minimum bits required for encoding is taken into account. This discipline is used in data mining to get an understandable prioritization of data sets with complex structures.

2. DATA MINING

Data mining is the process of discovering useful information from large sets of data. Data mining uses mathematical analysis to find out patterns and trends that exist in data. These patterns and trends can be collected and defined as a data mining model. Mining models can be applied to specific scenarios, such as finding Sequences, forecasting, grouping.

The area of data, data mining tasks, and data mining approaches faces many challenging research matters in data mining. The development of efficient and effective data mining methods and systems, the construction of interactive and integrated data mining environments, the design of data mining languages, and the application of data mining techniques to solve large application problems are important jobs for data mining researchers and data mining system and application developers. Several well-established statistical methods have been introduced for data analysis, such as regression, generalized linear models, analysis of variance, mixed-effect models, factor analysis, discriminant analysis, time-series analysis, survival analysis, and quality control. Researchers have been attempting to build theoretical foundations for data mining. Several interesting proposals have appeared, based on data reduction, data compression, pattern discovery, probability theory, microeconomic theory, and inductive databases. There are many data mining systems and research prototypes to choose from. When selecting a data mining product that is appropriate for one's task, it is important to consider various features of data mining systems from a multidimensional point of view. These include data types, system issues, data sources, data mining functions and methodologies, the tight coupling of the data mining system with a database or data warehouse system, scalability, visualization tools, and data mining query language and graphical user interfaces. Many customized data mining tools have been developed for domain-specific applications, including finance, the retail industry, telecommunications, bioinformatics, intrusion detection, and other science, engineering, and government data analysis. Such Scheme integrates domain-specific knowledge with data analysis methods and provides mission-specific data mining solutions. The Proposed research work acts as a root for new works and assessment proof of the current work.

3. PRINCIPLES OF KNOWLEDGE DISCOVERY

Knowledge discovery is the phenomenon of finding previously unknown patterns and designs from a big volume of data sets and converting the obtained patterns into understandable and applicable knowledge information. This domain of Knowledge discovery consists of many processes. These processes can take place at various stages, which form the basic rules of Knowledge Discovery domain. These processes are

Data Orientation: gather all the required data and building up one single accessible repository.

Data cleansing: Data is processed, analyzed and processed for better procedural treatment.

Data Selection: Selecting the required data from the given data sets for obtaining pattern.

Pattern Identification: The data sets are treated to get unknown patterns and designs

4. KNOWLEDGE DISCOVERY PROCESS

Knowledge discovery is the process of finding knowledge in the given data-sets irrespective of their characteristics and size attributes. The process of understanding and extracting the pattern from the given databases comprises of many steps. It is clearly illustrated in the following figure

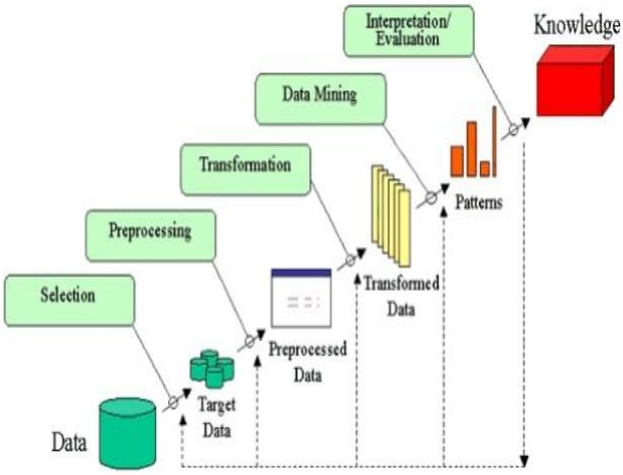

Figure: Knowledge Discovery Process

When a database is selected for data analysis, five main aspects should be considered. They are the factors to which the database belongs, Knowledge Prerequisite that is needed to understand the database, Application Knowledge required to obtain the required characteristics, Objective that has to be fulfilled after data pattern extraction and the level of attainment that is achieved after the pattern is discovered from the database. For the data pattern extraction process to be successful, we have to select the concerned dataset and create the desired variables that are required for correct or matching and analysis of the database. If the concerned variable is not properly done into the database, then it will result in biased data pattern output. After the concerned variable is fixed, the data-set is cleansed and is made to go into the pre-processing process. In the data cleansing stage, the data is removed from all the noises, incompleteness. These data-noises should be counted by collecting needed information for future enhancement of the removal process. Once the noises are removed after proper accounting made for them, data is reduced and are projected based on the objective that has to be fulfilled after data pattern extraction and the level of attainment that is gain after the pattern is discovered from the database. Once the data projection is made, the method of extraction of the data should be selected. The method that is to be employed for data extraction has to be carefully chosen because the method that is employed determines the level of uniqueness in data extraction process.

5. FUTURE WORK

Data mining is defined as the phenomenon of discovering patterns from huge amount of data. The process should be automatic or (more usually) semi-automatic. The patterns

discovered should be relevant in that they lead to some advantage, mostly an economic one. The data is invariably present in substantial quantities. And how are the patterns expressed? Relevant patterns allow us to make nontrivial predictions on new data. There are two extremes for the expression of a pattern: as a black box whose innards are effectively incomprehensible, and as a transparent box whose construction reveals the structure of the pattern. Both are making good predictions assumptions. The difference is whether or not the patterns that are mined are represented in terms of a structure that can be examined, reasoned about, and used to inform future decisions. Such patterns are structural because they capture the decision structure in an explicit way. In other words, they help to elaborate something about the data. The proposed framework, by itself, has a versatile ground work both literature wise and procedural wise. This framework was formulated by studying a lot of domain related literature works presented in various conference and journals. The proposed framework can be applied to various research studies. This framework can be used to aware the students in the domain of Data Mining and Knowledge Discovery. The future work will include the more detailed study of the related areas with a clear definition based on new analytical techniques that must be employed when it comes to analyzing data from inter disciplinary areas.

6. REFERENCES

[1] Pushpal Desai, Knowledge Discovery From Vehicle E-Governance Data Using Data Warehousing And Data Mining. International Journal of Information Technology & Management Information System (IJITMIS), 5 (2), 2014, pp. 40–50.
[2] Hill, S., Benton, A., Ungar, L., Macskassy, S., Chung, A. and Holmes, J.H., 2016. A Cluster-based Method for Isolating Influence on Twitter.
[3] Hill, S., Benton, A., Ungar, L., Macskassy, S., Chung, A. and Holmes, J.H., 2016. A Cluster-based Method for Isolating Influence on Twitter.
[4] Lorenzetti, C., Maguitman, A., Leake, D., Menczer, F. and Reichherzer, T., 2016. Mining for Topics to Suggest Knowledge Model Extensions. ACM Transactions on Knowledge Discovery from Data (TKDD), 11(2), p.23.
[5] Al-Hamidi, A., Lu, S. and Al-Salhi, Y., 2016. An enhanced frequent pattern growth based on MapReduce for mining association rules. International Journal of Data Mining & Knowledge Management Process (IJDKP) 2016, 6(2), pp.19-28
[6] Tayal, K. and Ravi, V., 2016, August. Particle Swarm Optimization Trained Class Association Rule Mining: Application to Phishing Detection. In Proceedings of the International Conference on Informatics and Analytics (p. 13). ACM.
[7] T. Venkatesan, T. Ramkumar and K. Saravanan, 2017, Mining Big Data: Towards a Machine Learning Framework Based on Collaborative Filtering. International Journal of Control Theory and Applications.
[8] Narang, S.K., Kumar, S. and Verma, V., 2017. Knowledge Discovery From Massive Data Streams. In Web Semantics for Textual and Visual Information Retrieval (pp. 109-143). IGI Global.
[9] Mathews, D., 2016. Data Mining and Machine Learning Algorithms for Workers' Compensation Early Severity Prediction (Doctoral dissertation, Middle Tennessee State University).
[10] Dasgupta, H., 2017. Data Mining and Statistics. Handbook of Research on Advanced Data Mining Techniques and Applications for Business Intelligence, p.15.
[11] Ramasubramanian, P., K. Iyakutti, and P. Thangavelu. "Enhanced data mining analysis in a higher educational system using rough set theory."African Journal of Mathematics and Computer Science Research 2, no. 9 (2009): 184-188.
[12] Ben-David, A., 2016. What does the Web remember of its deleted past? An archival reconstruction of the former Yugoslav top-level domain. new media & society, p.1461444816643790.
[13] Amani, F.A. and Fadlallah, A.M., 2017. Data mining applications in accounting: A review of the literature and organizing framework. International Journal of Accounting Information Systems, 24, pp.32-58
[14] Thakuriah, P.V., Tilahun, N.Y. and Zellner, M., 2017. Big data and urban informatics: innovations and challenges to urban planning and knowledge discovery. In Seeing Cities Through Big Data (pp. 11-45). Springer International Publishing.
[15] Wilson, S.J., 2017. Data representation for time series data mining: time domain approaches. Wiley Interdisciplinary Reviews: Computational Statistics, 9(1).