A SURVEY OF DATA MINING AND KNOWLEDGE DISCOVERY PROCESS MODEL AND ITS APPLICATIONS IN DATABASE

Nidhi Sharma* & Dr. Akash Saxena**
*Research Scholar, Sunrise University, Alwar.
**Supervisor of Sunrise University, Alwar.

Received: May 29, 2018 Accepted: July 01, 2018

ABSTRACT *Knowledge discovery and data mining have become areas of growing significance because of the recent increasing demand for KDD techniques, including those used in machine learning, databases, statistics, knowledge acquisition, data visualization, and high performance computing. The motive of mining is to find a new generation of computational theories and tools to assist humans in extracting useful information (knowledge) from the rapidly growing volumes of digital data. This article provides real-world applications, specific data mining techniques, challenges involved knowledge discovery. This paper also discusses relation between Knowledge and Data Mining, and Knowledge Discovery in Database.This survey presents a historical overview, description and future directions concerning a standard for a Knowledge Discovery and Data Mining process model. It presents a motivation for use and a comprehensive comparison of several leading process models, and discusses their applications to both academic and industrial problems. The main goal of this review is the consolidation of the research in this area. The survey also proposes to enhance existing models by embedding other current standards to enable automation and interoperability of the entire process.*

Keywords: *Knowledge discovery in databases, Data mining, Analysis, Information, Data mining applications, Knowledge management.*

INTRODUCTION: In information era, knowledge is becoming a crucial organizational resource that provides competitive advantage and giving rise to knowledge management (KM) initiatives. Many organizations have collected and stored vast amount of data. However, they are unable to discover valuable information hidden in the data by transforming these data into valuable and useful knowledge [2]. Managing knowledge resources can be a challenge. Many organizations are employing information technology in knowledge management to aid creation, sharing, integration, and distribution of knowledge.The rapid evolution of huge amount of data has lead to need for automated extraction of useful knowledge from such a huge dramatic pace. To extract this useful information some technique is required to focus on aspects of finding understandable patterns that can be interpreted as useful or interesting knowledge. So, as shown in Fig. 1, Data Mining is iterative and interactive process of discovering valid, novel, useful, and understandable knowledge (patterns, models, rules etc.) in Massive databases.

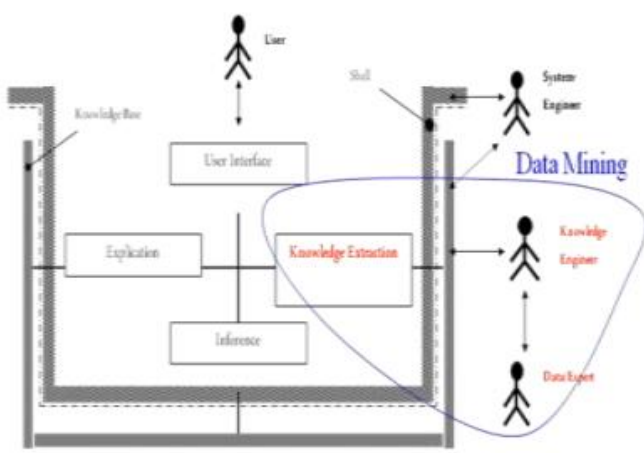

Fig. 1: Data Mining

Knowledge management is a process of data usage [6]. The basis of data mining is a process of using tools to extract useful knowledge from large datasets; data mining is an essential part of knowledge management [6].Wang & Wang (2008) point that data mining can be useful for KM in two main manners: (i) to share common knowledge of business intelligence (BI) context among data miners and (ii) to use data mining as a tool to extend human knowledge. Thus, data mining tools could help organizations to discover the hidden knowledge in the enormous amount of data. As a part of data mining research, this paper focuses on

surveying data mining applications in knowledge management through a literature review of articles from 2007 to 2012. The reason for reviewing research article this period is that data mining has emerged in KM research theme since 2006 and it plays important roles as a link between business intelligence and knowledge management.

These techniques and tools are the subject of the emerging field of knowledge discovery in databases (KDD) to discover strategic information hidden in very large databases.

1) Provide an overview of existing techniques that can be used for extracting of useful information from databases.

2) Provide a feature classification scheme that identifies important features to study knowledge discovery and data mining software tools.

3) Investigate existing knowledge discovery and data mining software tools using the proposed feature classification scheme. These tools may be either commercial packages available for purchasing, or research prototypes developed at various universities.

4) Practical application issues of KDD and enumerate challenges for future research and development.

REVIEW OF LITERATURE: Keeney (2006) in his examination expressed that assurance of business and data mining (specialized) targets is an imperative part of the KDDM process. This segment speaks to the beginning stage of the KDDM process. Given this reality, it is straightforward that disgraceful plan of destinations can prompt risking the whole KDDM project. Data mining literature and process models perceive the essentialness of this part, yet don't give any ways to deal with actualizing it. We recognize a few methodologies proposed in the literature that can be utilized. In the first place, we talk about esteem centered reasoning or VFT proposed as methods for detailing destinations and objectives. Second, we talk about SMART approach for figuring destinations that is regularly suggested in the professional literature.

Berry and Linoff (2007) expressed that Automatic cluster detection is utilized for finding significant patterns in data. Clustering gives an approach to find out about the structure of complex data. Once the correct clusters have been characterized, usually conceivable to discover basic patterns inside each cluster, as examine the accompanying qualities of automatic cluster detection, in clustering, there is no pre-characterized data and no refinement amongst autonomous and subordinate factors. In a more extensive sense, nonetheless, clustering can be a coordinated movement since clusters are looked for some business reason. In promoting, clusters shaped for a business reason for existing are typically called "segments", and client division is a well-known utilization of clustering. Automatic cluster detection is a data mining method that is seldom utilized as a part of disengagement since discovering clusters isn't regularly an end in itself.

Osei-Bryson (2004) recommends a multi criteria decision making way to deal with direct choice of the best decision tree from an extensive arrangement of decision trees. The recommended approach depicts the sorts of criteria that could be utilized for assessing the execution of decision trees and utilizing them in a multi-criteria decision making structure to help determination of the best mode.

Linstone and Turoff (2005) examined that Delphi might be described as a strategy for organizing a gathering correspondence process so the processes viable in permitting a gathering of people, overall, to manage a mind boggling issue. It is ended up being a famous device in IS examine in distinguishing and organizing issues for administrative decision making. Delphi is likewise applicable to the assessment venture of the KDDM process, as chose assessment criteria should be organized before data mining models can be chosen. Data mining issue composes are for the most part arranged into grouping, estimation, prediction, affiliation principles, clustering and perception. A somewhat unique plan and arrange issues in view of the demonstrating expectation as (a) displaying to comprehend, (b) demonstrating to order, and (c) displaying to anticipate.

Schenkerman (2007) explored that AHP prescribes disintegrating an issue into an arrangement of elements, doling out numerical weights or needs to those elements, and looking at changed choices agreeing based on their scores on the picked set of elements. These different options would then be able to be rank ordered to make a determination. One of the main qualities of AHP is that it can catch both subjective and also target assessment criteria. While AHP has been over a wide assortment of decision circumstances, it isn't without feedback. Pundits of AHP have indicated inconsistency of results inferable from utilization of subjective scales, rank inversions, and Inducement of Nonexistent Order and so forth. Verbal confrontations between the faultfinders and advocates have likewise been exhibited in the literature. Notwithstanding, AHP keeps on being utilized as a prevalent decision making instruments by specialists and academicians, It has additionally been joined in type of the business programming Expert Choice.

Inmon (2012) expressed in an exploration that a data warehouse (DW) is a gathering of coordinated, subject-situated databases intended to help the DSS (decision bolster) work, where every unit of data is nonvolatile and applicable to some minute in time. The Data Warehouse comprises of operational data stores and data shops. The operational data store (ODS) is the most widely recognized segment of the DW condition. Its essential everyday capacity is to store the data for a solitary, particular arrangement of operational applications. The data shop is regularly seen as an approach to pick up passage into the domain of data warehouses and to commit all errors on a littler scale.

DATA MINING: Data mining is an essential step in the knowledge discovery in databases (KDD) process that produces useful patterns or models from data (Figure 2) [7]. The terms of KDD and data mining are different. KDD refers to the overall process of discovering useful knowledge from data. Data mining refers to discover new patterns from a wealth of data in databases by focusing on the algorithms to extract useful knowledge [7].

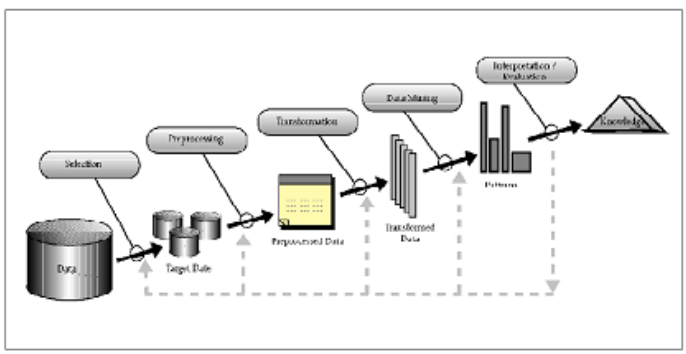

Figure 2 Data Mining and the KDD Process

Based on figure 2, KDD process consists of iterative sequence methods as follows [7, 9]:

1. Selection: Selecting data relevant to the analysis task from the database
2. Preprocessing: Removing noise and inconsistent data; combining multiple data sources
3. Transformation: Transforming data into appropriate forms to perform data mining
4. Data mining: Choosing a data mining algorithm which is appropriate to pattern in the data; extracting data patterns
5. Interpretation/Evaluation: Interpreting the patterns into knowledge by removing redu ndant or irrelevant patterns; translating the useful patterns into terms that human understandable

KNOWLEDGE MANAGEMENT: There are various concepts of knowledge management. In this paper we use the definition of knowledge management by McInerney (2002): "Knowledge management (KM) is an effort to increase useful knowledge within the organization. Ways to do this include encouraging communication, offering opportunities to learn, and promoting the sharing of appropriate knowledge artifacts" This definition emphasizes the interaction aspect of knowledge management and organizational learning. Knowledge management process focuses on knowledge flows and the process of creation, sharing, and distributing knowledge (Figure 3) [5]. Each of knowledge units of capture and creation, sharing and dissemination, and acquisition and application can be facilitated by information technology.

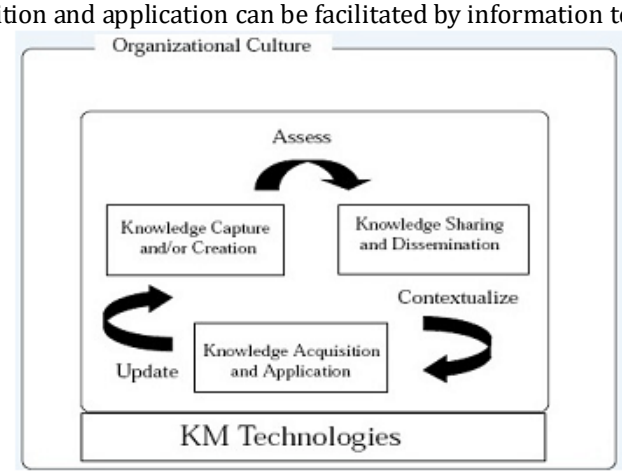

Figure 3 KM Technologies Integrated KM Cycle

As technologies play an important role in KM, technologies stand to be a necessary tool for KM usage [1]. Thus, KM requires technologies to facilitate communication, collaboration, and content for better knowledge capture, sharing, dissemination, and application [5].

DATA MINING STEP OF THE KDD PROCESS: Data mining process is used to extract information from a data set and transform it into an understandable structure for further use as shown in Fig. 4. The data mining component of the KDD process often involves repeated iterative application of particular data mining methods. It is achieved by using application domain like prior knowledge, user goals etc. to create target dataset that will be used in data miningalgorithms that interpret, evaluate and visualize patterns and manage discovered knowledge.

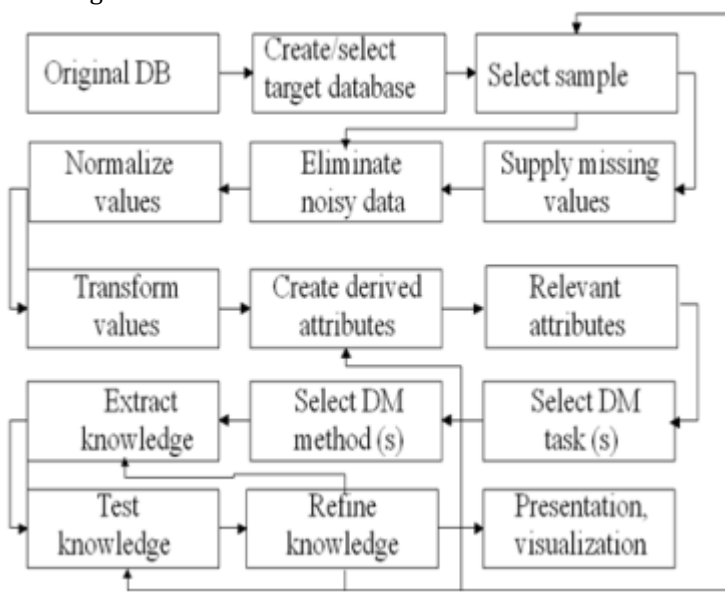

Fig. 4: Data mining process

The knowledge discovery goals are defined by the intended use of the system. We can distinguish two types of goals: Verification, where the system is limited to verifying the user's hypothesis, and Discovery, where the system autonomously finds new patterns. In this paper we are primarily concerned with discovery-oriented data mining. The goals of prediction and description are achieved via the following primary data mining methods:

1) Classification: learning a function that maps (classifies) a data item into one of several predefined classes.

2) Regression: learning a function which maps a data item to a real-valued prediction variable and the discovery of functional relationships between variables.

3) Clustering: identifying a finite set of categories or clusters to describe the data. Closely related to clustering is the method of probability density estimation which consists of techniques for estimating from data the joint multi-variant probability density function of all of the variables/fields in the database.

4) Summarization: finding a compact description for a subset of data, e.g., the derivation of summary or association rules and the use of multivariate visualization techniques.

5) Dependency Modeling: finding a model which describes significant dependencies between variables (e.g., learning of belief networks).

6) Change and Deviation Detection: discovering the most significant changes in the data from previously measured or normative values.

CONCLUSION: In this paper, we have discussed detail study of data mining with various studies like tasks, techniques, applications and challenging issues. A primary aim is to clarify the relation between knowledge discovery and data mining. We provided an overview of the KDD process and basic data mining methods. The implementation of data mining techniques will allow users to retrieve meaningful information from virtually integrated data. These techniques provide variety of applications for industries like retail, telecommunication, Biomedical etc. These tools predict future trends and behaviors, allowing business to make proactive and present knowledge in the form which is easily understood to human. The focus has been

given on fundamental methods for conducting data mining. The methods include natural language processing and information extraction. A brief review on application domains has been presented. The purpose of this section is to give an overview to a reader on how text mining systems can be used in real life. The paper also addressed the most challenging issue in developing data mining systems.

REFERENCES:

1. Keeney, R. L. (2006). "Value-Focused Thinking: Identifying Decision Opportunities and Creating Alternatives." European Journal of Operations Research 92: 537- 549.
2. Berry, M. and G. Linoff (2007). Data Mining Techniques for Marketing, Sales and Customer Support, John Wiley and Sons.
3. Osei-Bryson, K.-M. (2004). "Evaluation of Decision Trees." Computers and Operations Research 31: 1933-1945
4. Linstone, H. A. and M. Turoff (2005). The Delphi Method: Techniques and Applications.
5. Schenkerman, S. (2007). "Inducement of nonexistent order by the analytic hierarchy process." Decision Sciences 28(2): 475-482.
6. Inmon, W. H. (2012). Building the Data Warehouse.New York, Wiley.
7. Gorunescu, F. (2011). Data Mining: Concepts, Models, and Techniques. India: Springer.
8. Han, J. &Kamber, M. (2012). Data Mining: Concepts and Techniques. 3rd.ed. Boston: Morgan Kaufmann Publishers.
9. Hwang, H.G., Chang, I.C., Chen, F.J. & Wu, S.Y. (2008). Investigation of the application of KMS for diseases classifications: A study in a Taiwanese hospital. Expert Systems with Applications, 34(1), 725-733. doi:10.1016/j.eswa.2006.10.018
10. Lavrac, N., Bohanec, M., Pur, A., Cestnik, B., Debeljak, M. &Kobler, A. (2007).Data mining and visualization for decision support and modeling of public health-care resources.Journal of Biomedical Informatics, 40, 438-447. doi:10.1016/j.jbi.2006.10.003