Comprehensive Analysis of Data Mining Techniques and Trends for Knowledge Management System

Dr. Anubhav Kumar[1], Dr. Arvind K Sharma[2]

[1]Dept. of CSE, Sri Sai College of Engineering & Technology, Badhani

[2]Dept. of CSI, University of Kota, Rajasthan, India

(*E-mail:* dr.anubhavkumar@gmail.com, drarvindkumarsharma@gmail.com)

***Abstract*— *Data mining is often used during the knowledge discovery process and is one of the most important sub fields in knowledge management. Knowledge management is the system and managerial approach to the gathering, management, use, analysis, sharing, and discovery of knowledge in an organization Due to the importance of extracting knowledge/information from the large data repositories, data mining has become an essential component in various fields of human life including business, education, medical, scientific etc. This paper explores the techniques and trends of data mining which developed to support knowledge management system. Finally, the applications of data mining techniques in the process of knowledge management are summarized and discussed.*

Keywords: Data mining, Knowledge management,

I. INTRODUCTION

Database technology since the mid-1980 has been characterized by the popular global adaptation of relational model and drastic change of research and development activities on new and powerful database systems. These employ advanced data model. The exponential growth of computer hardware and system software technology in the past three decades has led to large supplies of powerful and cost effective computers, data collection equipment and storage media. This technology provides a great boost to the database and information industry and makes a huge number of databases and information repositories available for transaction management information retrieval and data analysis. Data mining is simply an essential step in the process of knowledge discovery from the databases. Data mining uses a combination of an explicit knowledge base, sophisticated analytical skills, and domain knowledge to uncover hidden trends and patterns. These trends and patterns constitute the basis of predictive models that enable analysts to produce new observations from existing data [8, 9]. 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 [10]. 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. Knowledge management is a process of data usage [11]. 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 [11].

The rest of this paper is organised as follows: Section 2 presents an introduction to data mining. Section 3 describes Knowledge management system. Section 4 contains the different useful data mining techniques. Section 5 discusses the techniques, trends and data mining areas. Section 6 concludes the paper while references are mentioned in the last.

II. DATA MINING

Data mining is an essential step in the knowledge discovery in databases (KDD) process that produces useful patterns or models from data which shown in fig. 1. 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 [1]. Following figure 1 refers the Data mining is one among the most important core step in the Knowledge Discovery in Databases (KDD) process. It can be considered as a heart of the KDD process.

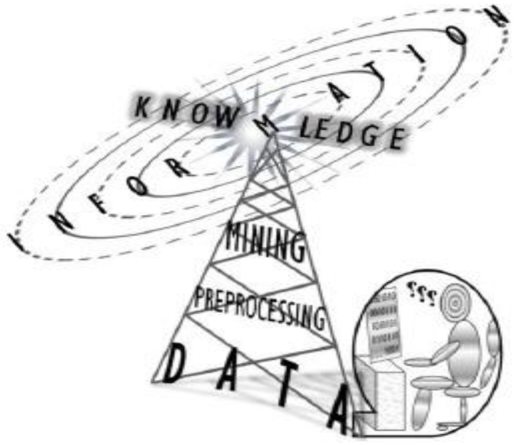

Fig.1: Data Mining is a Heart of KDD Process

The KDD process contains selecting the data needed for data mining process and may be obtained from many different and heterogeneous data sources. Preprocessing includes finding incorrect or missing data [6]. There may be many different activities performed at this time. Erroneous data may be corrected or removed, whereas missing data must be supplied. Preprocessing also include: removal of noise or outliers, collecting necessary information to model or account for noise, accounting for time sequence information and known changes. Transformation is converting the data into a common format for processing. Some data may be encoded or transformed into more usable format. Data reduction, dimensionality reduction (e.g. feature selection i.e. attribute subset selection, heuristic method etc) & data transformation method (e.g. sampling, aggregation, generalization etc) may be used to reduce the number of possible data values being considered. Data Mining is the task being performed, to generate the desired result. Interpretation/Evaluation is how the data mining results are presented to the users which are extremely important because the usefulness of the result is dependent on it. Various visualization and GUI strategies are used at this step. Knowledge discovery as a process consists of an iterative sequence of the following steps:

i) Data Cleansing (to remove noise and inconsistent data)

ii) Data integration (where multiple data sources may be combined)

iii) Data Selection (Where data relevant to the analysis task are retrieved from the database)

iv) Data transformation (Where data are transformed or consolidated into forms appropriate for mining by performing summery or aggregation operations for instance

v) Data mining (an essential process where intelligent methods are applied in order to extract data patterns)

vi) Pattern evolution (to identify the truly interesting patterns representing knowledge based on some interestingness measures)

vii) Knowledge presentation (where visualization and knowledge representation techniques are used to present the mined knowledge to the user) [2].

III. KNOWLEDGE MANAGEMENT SYSTEM

A. Definition of 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. A knowledge management system (KMS) framework can be visualized as a system shown in fig. 2. In Knowledge Management System, data and rules enter into the system as inputs, knowledge extraction is undertaken based on the input rules, the extracted knowledge is managed and knowledge based services are provided to the stake holders[5].

A Knowledge Management System can be split into following sub components:

- **Repositories**- These hold explicated formal & informal knowledge and the rules associated with them for collection, refining, managing, validating, maintaining, interpreting and distributing content.
- **Collaborative Platforms**- These support distributed work and incorporate pointers, skills databases, expert locators and informal communication channels.
- **Networks**- Networks support communications and conversion. They include broad bands, leased lines, intranets, extranets etc.
- **Culture**- Culture enablers that encourage sharing and use.

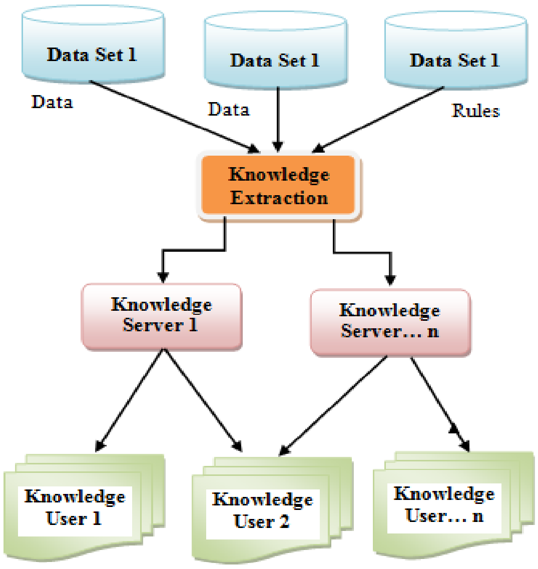

Fig. 2: Knowledge Management System

IV. DATA MINING TECHNIQUES

This section will focus on the conventional techniques utilized by data analysts to implement the data mining algorithms. Data mining identifies facts or suggests conclusions based on shifting through the data to discover either patterns or anomalies. Data mining comprises the following techniques[3,4]:

A. Classification: The most common technique used in mining is Classification [12]. Classification, as the name suggests, allows the user to classify large populous data into a model which sorts them into a predefined set of classes. Fraud detection, classifying patients from primary health care centers to specialists, and credit risk applications are a few ways in which classification is implemented [13]. Classification process often employs supervised learning and classification, and is mostly used for predictive modeling. Some of the popular algorithmic models employed in classification are decision trees, neural networks, Bayesian classification, Support Vector Machines (SVM) and classification based on association.

B. Clustering:

Clustering is another DMT which has gained popularity among the mining community. It involves identifying clusters and grouping similar objects together in each cluster. While classification is mentioned to have employed supervised learning, clustering process mainly uses unsupervised learning method (some clustering models use both) [14]. Analyzing the similarity in organizational behavior, financial trends and clustering homes based on energy consumption are a few algorithms used in this technique. Even though researchers have mainly focused on evaluating and implementing Partitioned (K-means) algorithms [15,16], other clustering methods include Hierarchical (CURE, BIRCH), Grid – based (STING, WaveCluster), Model-based (Cobweb) [17], and Density based (DBSCAN) [18].

C. Regression

The term regression is defined as an analyzing or measuring the relation between a dependent variable and one or more independent variable. Regression techniques can be categorized in two categories such as: Linear regression and Logistic regression which are shown in fig. 3.

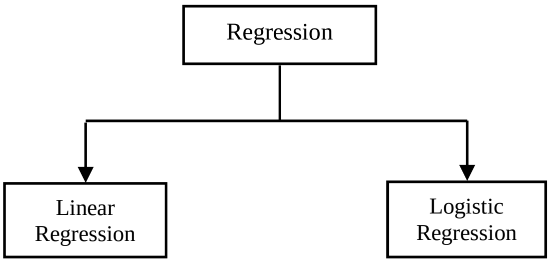

Fig. 3: Classification of Regression

C.1. Linear Regression

Linear Regression was historically the earliest predictive method and is based on the relationship between input variables and the output variable. Linear regression is a simple technique suitable for numeric prediction that is frequently used in statistical application. The idea is to find the amount of how much each of the attributes $\mathbf{a_1}$, $\mathbf{a_2}$,..., $\mathbf{a_k}$ in a data set contributes to the target value x . Each attribute is assigned a factor $\boldsymbol{w_i}$ and one extra factor is used to constitute the base level of the predicted attribute.[19]

$$\boldsymbol{x = w + w_1a_1 + w_2a_2 + ... + w_ka_k}$$

The aim of linear regression is to find optimal weights for the training instances by minimizing the error between the real and the predicted values. As long as the data set contains more instances than attributes this is easily done using the least square method [20]. Linear regression is quite intuitive and easily understood but the downside is that it handles non-numerical attributes poorly and that it can't handle more complex nonlinear problems [21].

C.2 Logistic Regression

Logistic regression is a generalization of linear regression [22]. Basically it is used for estimating binary or multi-class dependent variables and the response variable is discrete, it cannot be modelled directly by linear regression i.e. discrete variable changed into continuous value. It also provides the difference in the percentage of dependent variable and provides the rank of individual variable according to its importance. Thus the main aim of Logistic regression is to determine the result of each variable correctly. Logistic regression is also known as nominal regression[19]. It is a statistical technique for classifying records based on values of input fields. It is analogous to linear regression but takes a categorical target field instead of a numeric one. Both binomial models (for targets with two discrete categories) and multinomial models (for targets with more than two categories) are supported. It works by building a set of equations that relate the input field values to the probabilities associated with each of the output field categories. Once the model is generated, it can be used to estimate probabilities for new data. For each record, a probability of membership is computed for each possible output category. The target category with the highest probability is assigned as the predicted output value for that record. Linear regression models are often quite accurate. They can handle symbolic and numeric input fields. They can give predicted probabilities for all target categories. Logistic models are most effective when group membership is a truly categorical field. Logistic regression is related to some other statistical analysis techniques but it offers more flexibility and robustness [23,24]. It does not assume linear relationship between the input and output variables, nor normal distribution and equal variance within input variables.

D. Association Rules Mining: Association rules mining is used to search correlation relationships among a large set of data items or variables. The association rules can be seen as the identification of actions or facts that, being initially independent, they happen in a combined or associate way. The considered facts can be characteristics or behaviors observed in the individuals. A typical example of association rules mining is the market basket analysis[25] i.e. a supermarket might gather data on customer purchasing habits. Using association rule learning, the supermarket can evaluate which products are frequently bought together and use this information for marketing purposes. This is sometimes referred to as Market Basket Analysis.

V. COMPREHENSIVE ANALYSIS OF DATA MINING TECHNIQUES & TRENDS

The field of data mining has been growing due to its enormous success in terms of broad-ranging application achievements and scientific progress, understanding. Many data mining applications have been successfully implemented in several domains such as- health care, finance, retail, telecommunication, fraud detection and risk analysis etc. The ever increasing complexities in various fields and improvements in technology have posed new challenges to data mining. These various challenges contains different data formats, data from disparate locations, advances in computation and networking resources, research and scientific fields, ever growing business challenges etc. Advancements in data mining with various integrations and implications of methods and techniques have shaped the present data mining applications to handle the various challenges, the current trends and techniques of data mining applications have been presented in this section. The table-1 shows several currently employed data mining techniques and algorithms to mine the various data formats in different application areas especially for Knowledge Management domains. The techniques, trends and application areas of data mining are summarized in table-1 and table-2 (Please refer to Appendix-A). Besides it, a comprehensive analysis of several data mining trends from past to future in context of Knowledge management domains is shown in table-2. It includes different techniques, data formats and computing resources used in different applications in past, current and future. (Please refer to Appendix-B).

VI. CONCLUSION

In this paper we have been tried to briefly review the several data mining techniques and trends from its inception to the future in context of knowledge management domains. A comprehensive analysis of several data mining trends and data formats from past to the future in context of Knowledge management system. It is shown that data mining becoming increasingly popular in both the private and public sectors. Several Industries such as banking, insurance, medicine, retailing, education sector commonly use data mining to reduce costs, enhance research and increase sales. Thus, Data mining with Knowledge management will be more and more useful in future. This paper shall definitely be helpful to the researchers to keep their focus on different data mining issues in context of Knowledge management.

REFERENCES

[1] Fayyad, U., Piatetsky-Shapiro et. al, From Data Mining to Knowledge Discovery in Databases. AI Magazine, 17(3), 37-54,1996.

[2] Peter P. Wakabi-Waiswa Venansius Baryamureeba, Extraction Of Interesting Association Rules Using Genetic Algorithms International Journal of Computing and ICT Research, 2(1), 2008.

[3] Introduction to Data Mining and Knowledge Discovery, Third Edition ISBN: 1-892095-02-5, Two Crows Corporation, 10500 Falls Road, Potomac, MD 20854 (U.S.A.), 1999.

[4] Campos, M. M., et al., “Data-Centric Automated Data Mining”, www.oracle.com/technology/ products/bi/odm/ pdf/automated_data_mining_paper_1205.pdf.

[5] Abdullah, A., Brobst, S., M.Umer M. (2004). "The case for an agri data ware house: Enabling analytical exploration of integrated agricultural data". Proc. of IASTED International Conference on Databases and Applications, Austria.

[6] Arvind K. Sharma and P.C. Gupta, “Exploration of efficient methodologies for the Improvement in web mining techniques: A survey”, International Journal of Research in IT & Management (ISSN 2231-4334) Vol.1, Issue 3, July 2011

[7] [Online] http://www.cs.waikato.ac.nz/ml/weka

[8] Kidwell, J. J, Vander Linde, K. M., Johnson, S. L, “Knowledge Management Practices in Higher Education”, Educause Quarterly, 4/2000, pp.28-33, 2000

[9] Kittler, R., Wang, W., “The Emerging Role for Data Mining”, Solid State Technology, 42(11), pp. 45-58, November 1999

[10] [10] Berson, A., Smith, S.J. &Thearling, K. (1999). Building Data Mining Applications for CRM. New York: McGraw-Hill.

[11] [12] Dawei, J. (2011). The Application of Date Mining in Knowledge Management.2011 International Conference on Management of e-Commerce and e-Government, IEEE Computer Society, 7-9. doi: 10.1109/ICMeCG.2011.58

[12] B. M. Ramageri, "DATA MINING TECHNIQUES AND APPLICATIONS," Indian Journal of Computer Science and Engineering, Vol. 1, No. 4, pp. 301-305, 2011.

[13] R. Petre, "Data Mining Solutions for the Business Environment," Database Systems Journal, Vol. 4, pp. 21-28, 2013.

[14] L. Rokach and O. Maimon, "Clustering Methods," in The Data Mining and Knowledge Discovery Handbook, New York, Springer US, 2006, pp. 321--352.

[15] [N. Sharma, A. Bajpai and R. Litoriya, "Comparison the various clustering algorithms of weka," International Journal of Emerging Technology and Advanced Engineering, vol. 2, no. 5, pp. 73-80, 2012.

[16] S. Kumar and N. , "K-Mean Evaluation in Weka Tool and Modifying It using Standard Score Method," International Journal on Recent and Innovation Trends in Computing and Communication, vol. 2, no. 9, p. 2704 – 2706, 2014.

[17] J. Han and M. Kamber, Data Mining - Concepts and Techniques, 2nd Edition ed., San Fransisco: Elsevier, 2008.

[18] P. Shrivastava and H. Gupta, "A Review of Density-Based clustering in Spatial Data," International Journal of Advanced Computer Research, vol. 2, no. 5, pp. 200-202, 2012.

[19] Arvind K. Sharma, P.C. Gupta, “Performance Evaluation Of Cpu For Regression Technique Using Data Mining Tool- Weka”, International Journal of Management, IT and Engineering, Vol.3, Issue-2, 2013

[20] I. H. Witten and E. Frank, “Data Mining: Practical Machine Learning Tools and Techniques”, Morgan Kaufmann, Amsterdam, 2005.

[21] M. H. Dunham. Data mining: Introductory and Advanced Topics, Prentice Hall/Pearson Education, Upper Saddle River, NJ, 2003.

[22] De Mantaras & Armengol E. (1998), “Machine learning from example: Inductive and Lazy methods”, Data & Knowledge Engineering 25: 99-123.

[23] B. Hamadicharef, et al., “Performance evaluation and fusion of methods for early detection of alzheimer disease”, In Proc. Int. Conf. BioMedical Engineering and Informatics BMEI 2008, vol. 1, pp. 347–351, May 2008.

[24] B. G. Tabachnick et al., “Using Multivariate Statistics”, 5th edition, Allyn & Bacon, Inc., Needham Heights, MA, USA, pp. 437-505, 2007

[25] Srivastava, J., Cooley, R., Deshpande, M., Tan, P. “Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data” SIGKDD Explorations, 2000

APPENDIX – A

Table-1: Data Mining Techniques and Trends to Mine Several Data Formats

Data Mining Type	Trends/Application Areas	Data Formats	Data Mining Techniques/Algorithms
Hypermedia data mining	Internet and Intranet Applications	Hyper Text Data	Classification and Clustering Techniques
Ubiquitous data mining	Applications of Mobile phones, PDA, Digital Cam etc.	Ubiquitous Data Traditional data mining techniques drawn from the Statistics and Machine Learning	Traditional data mining techniques drawn from the Statistics and Machine Learning
Multimedia data mining	Audio/Video Applications	Multimedia Data	Rule based decision tree classification algorithms
Spatial Data mining	Network, Remote Sensing and GIS applications.	Spatial Data	Spatial Clustering Techniques, Spatial OLAP
Time Series Data mining	Business and financial applications.	Time series data	Rule Induction algorithms

APPENDIX – B

Table-2: Comprehensive Analysis of Data Mining Techniques and Trends

Data Mining Trends	Techniques/ Algorithms Employed	Data Formats	Computing Resources	Application Areas
Past	Statistical, Machine Learning Techniques	Numerical data and structured data stored in traditional databases	Evolution of 4G PL and various related techniques	Business
Current	Statistical, Machine Learning, Artificial Intelligence, Pattern Reorganization Techniques	Heterogeneous data formats includes structured, semi structured and unstructured data	High speed networks, High end storage devices and Parallel, Distributed computing etc…	Business, Web, Medical diagnosis etc.
Future	Soft Computing techniques like Fuzzy logic, Neural Networks and Genetic Programming	Complex data objects includes high dimensional, high speed data streams, sequence, noise in the time series, graph, Multi instance objects, Multi represented	Multi-agent technologies and Cloud Computing	Business, Web, Medical diagnosis, Scientific and Research analysis fields (bio, remote sensing etc.), Social networking etc.

		objects and temporal data etc.		