International Journal of
Emerging Research in Management &Technology
ISSN: 2278-9359 (Volume-4, Issue-9)

Research Article | **September 2015**

Extracting Data Mining Applications through Depth-Analyzed Data

[1]Rohit Yadav, [2]Kratika Varshney, [3]Kapil Arora

[1, 2] Information Technology Department, Aligarh College of Engineering &Technology, Aligarh, Uttar Pradesh, India

[3] Computer Science & Engg. Department, Aligarh College of Engg & Technology, Aligarh, Uttar Pradesh, India

Abstract:

In this paper we have focused about on what data mining is and how we can use the data mining in the various fields. Data mining is a new powerful technology that helps business to focus on important information like future trends, decision making, customer choice etc. Data mining is the knowledge discovery process by analyzing the large volumes of data from various perspectives and summarizing it into useful information. Data mining is the process of extracting the information and patterns derived by the KDD process which helps in crucial decision-making. Data mining works with data warehouse and the whole process is dived into action plan to be performed on data: Selection, transformation, mining and results interpretation. In this paper we will discuss the variety of techniques, approaches and different areas of the research which are helpful and marked as the important field of data mining Technologies. In this paper we will discuss basics of educational data mining.

Keywords: Data Mining, Definition, Trends, KDD, DM Steps and Technique, Research and Application Challenges for KDD, Application of Data Mining in Various Fields.

I. INTRODUCTION

Data mining is the process of finding of hidden information from a huge amount of data. Data mining analyzing the data from different source and convert it into meaningful information**.** Data mining is the process of discovering insightful, interesting, and novel patterns, as well as descriptive, understandable, and predictive models from large-scale data. We emphasize the geometric and algebraic views, as well as the probabilistic interpretation of data. A target dataset is prepared before applying the data mining algorithm. The common source of data is the data warehouse. Pre–processing is needed to analyze the data sets before applying the data mining. Each and every day the human beings are using the vast data and these data are in the different fields .It may be in the form of documents, may be graphical formats ,may be the video, may be records (varying array) .As the data are available in the different formats so that the proper action to be taken. Not only to analyze these data but also take a good decision and maintain the data .As and when the customer will required the data should be retrieved from the database and make the better decision .This technique is actually we called as a data mining or Knowledge Hub or simply KDD(Knowledge Discovery Process).

II. DATA MINING DEFINITION

"Data Mining represents a process developed to examine large amounts of data routinely collected. The term also refers to a collection of tools used to perform the process. Data mining is used in most areas where data are collected-marketing, health, communications, etc."

Trrends:

Historical Trends of Data Mining: Data mining is useful in various disciplines, which includes database management systems (DBMS), Statistics, Artificial Intelligence (AI), and Machine Learning (ML). The era of data mining applications was conceived in the year1980 primarily by research-driven tools focused on single tasks. The early day's data mining trends are as under.

1. Data Trends: In initial days, data mining algorithms work best for numerical data collected from a single data base, and various data mining techniques have evolved for flat files, traditional and relational databases where the data is stored in tabular representation. Later on, with the confluence of Statistics and Machine Learning techniques, various algorithms evolved to mine the non numerical data and relational databases.
2. Computer Trends: The field of data mining has been greatly influenced by the development of fourth generation programming languages and various related computing techniques. In, early days of data mining most of the algorithms employed only statistical techniques. Later on they evolved with various computing techniques like AI, ML and Pattern Reorganization. Various data mining techniques (Induction, Compression and Approximation) and algorithms developed to mine the large volumes of heterogeneous data stored in the data warehouses.

Current 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. Various data mining applications have been successfully implemented in various domains like 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; the various challenges include different data formats, data from disparate locations, advances in computation and networking resources, research and scientific fields, ever growing business challenges etc.

1. Bio-informatics and Cure for Diseases: The most important application trend, deals with mining and interpretation of biological sequences and structures. Data mining tools are rapidly being used in finding genes regarding cure of diseases like Cancer and AIDS.
2. Fight against Terrorism: After 9-11 attacks, many countries imposed new laws against fighting terrorism. These laws allow intelligence agencies to effectively fight against terrorist organizations.USA launched Total Information Awareness program with the goal of creating a huge database of that consolidate all the information on population. Similar projects were also launched in European countries and rest of the world. This program faced several problems, a. The heterogeneity of database, the target database had to deal with text, audio, image and multimedia data. b. Second problem was scalability of algorithms. The execution time increases as size of data (which is huge). For example, 230 cameras were placed in London, to read number plates of vehicles. An estimated 40,000 vehicles pass camera every hour, in this way the camera must recognize 10 vehicles per second, which poses heavy loads on both hardware and software.
3. Web and Semantic Web: Web is the hottest trend now, but it is unstructured. Data mining is helping web to be organized, which is called Semantic web. The underlying technology is Resource Description Framework (RDF) which is used to describe resources. FOAF is also a supporting technology, heavily used in Face book and Orkut for tagging. But still there are issues like combining all RDF statements and dealing with erroneous RDF statements. Data mining technologies are serving a lot to make the web, a semantic web.
4. Business Trends: Today's business environment is more dynamic, so businesses must be able to react quicker, must be more profitable, and offer high quality services that ever before. Here, data mining serves as a fundamental technology in enabling customer's transactions more accurately, faster and meaningfully. Data mining techniques of classification, regression, and cluster analysis are used for in current business trends. Almost all of the current business data mining applications are based on the classification and prediction techniques for supporting business decisions, thus creating strong Business Intelligence (BI) system.

III. KNOWLEDGE DISCOVERY PROCESS (KDD)

Data mining and knowledge discovery in databases are related to each other and to other related fields such as machine learning, statistics, and databases. Data Mining is one of the steps in the
overall process of KDD that consists of collection and preprocessing of data, data mining, interpretation, evaluation of discovered knowledge and finally post processing. The KDD field's basic objective is to make data meaningful by developing methods and techniques of mining but problem being faced by the KDD process is to map huge and heterogeneous data into understandable, more abstract and useful form .
The phrase knowledge discovery in databases emphasizes that knowledge is the end product of a data-driven discovery. The data-mining step of KDD relies heavily on known techniques from machine learning, pattern recognition, and statistics to find patterns from data.
The data mining step of the KDD process: Data mining step of KDD Process involves iterations for particular data-mining methods in application. There are two types of goals:

1. Verification in which system is limited to verifying user's hypothesis and,
2. Discovery, in which system autonomously finds new patterns.

Data Mining Method for KDD: Primary goals of data mining in practice are prediction and description. In prediction some variables and fields in the database are used to predict unknown values of other variables of interest, and description helps in finding human-understandable patterns describing the data Weiss and Kulikowski in [33] proposed that, "Classification is learning a function that maps (classifies) a data item into one of several predefined classes". Apte and Hong in [34] suggested that classification methods of Data mining are used as part of knowledge discovery applications which includes classifying trends in financial markets, education and identifying objects of interest from large dataset of images. Regression is a predictive technique that maps data item to a prediction variable. Clustering is a descriptive task where we identify a finite set of categories or clusters to describe the data. E.g. identifying those students who are short of attendance and shown poor performance in sessionals. Cheeseman and Stutz suggested that examples of clustering applications in a knowledge discovery context include discovering similar groups. Summarization involves methods like calculating mean and standard deviations. There are some methods which involve deriving of abstract rules, visualization techniques, and the discovery of functional relationships between variables. Summarization techniques are often applied to interactive exploratory data analysis and automated report generation.

The Components of Data mining Algorithm:
One can identify three primary components in any DM algorithm:

1. Model representation
2. Model evaluation
3. Search.

A model representation is used to describe or extract patterns whereas Model-evaluation criteria are statements which help in meeting the goals of Knowledge Discovery Process using particular pattern or model. Predictive models are

judged by the prediction accuracy on some dataset and descriptive models are evaluated along the dimensions of predictive accuracy, novelty, utility and understandability of the model. Search method consists of two components:

1. Parameter search
2. Model search.

Once the model representation and the model-evaluation criteria are fixed, then Data Mining problem left with optimization of task on observational dataset.

Data Mining Steps:

1. According to IBM report, three main steps in DM are preparing the data, reducing the data and, finally, looking for useful information.
2. Predictive modeling uses inductive reasoning techniques and algorithms like neural networks.
3. Database segmentation use statistical clustering techniques to partition data into clusters.
4. Link analysis identifies useful associations between data.
5. Deviation detection detects and explains why certain records cannot be put into specific segments.
6. Fayyad et al, proposed following steps of Data Mining: Retrieving the data from a large database. Selecting the relevant subset to work with.
7. Deciding appropriate sampling system, transformations, cleaning the data and to deal with missing fields and records.
8. Fitting models to the pre-processed data.

Data Mining Techniques: There are different data mining techniques which are used to extract information from a data set and transform it into an understandable format for further use. Table I shows different Data Mining Techniques and their roles.

1. **Statistics** is a vital component in data selection, sampling, Data Mining, and knowledge evaluation. In data cleaning process, statistics offer the techniques to detect outliers to simplify data when necessary, and to estimate noise, it deals with missing data using estimation techniques.
2. **Classification and prediction** One of the most useful data mining techniques for e-learning is classification. Classification maps data into predefined group of classes. Classification is supervised learning approach because the classes are determined before examining the data. The prediction of student's performance with high accuracy is more beneficial for identifying low academic performance of the students at the beginning. Classification is the processing of finding a set of models which describe and distinguish data classes or concepts. The derived results may be represented in various forms, such as classification rules, decision trees, or neural networks. Models then can be used for predicting the class label of data objects. In many applications, there is need to predict some missing data values rather than class labels. E.g. Case when the predicted values are numerical data, and is often specifically referred to as prediction.
3. **Clustering** groups the data, which is not predefined and it can identify dense and sparse regions in object space. Clustering algorithm groups the data. Unlike classification and prediction, which analyze class labeled data objects, clustering analyses data objects without consulting a known class label. The class labels are not present in the training data and clustering can be used to generate such labels. Clusters of objects are formed so that objects within a cluster have high similarity in comparison to one another, but are very dissimilar to objects in other clusters. Each cluster formed can be viewed as a class of objects, from which rules can be derived. Application of clustering in education can help in finding academic trends, student's performance analysis in class.
4. **Association** rule mining is to find set of binary variables that occurs in the transaction database repeatedly. Apriority measures are the association rule mining algorithm. Association analysis is the discovery of association rules showing attribute value conditions that occur frequently together in a given set of data. The association rule A=>B shows those database tulles that satisfy the conditions in A as well as in B.

IV. RESEARCH AND APPLICATION CHALLENGES IN KDD

1. Larger databases: There are databases with hundreds of fields, tables; millions of records and to derive some useful information from it is itself a challenge. Scientist suggested methods for dealing with large data volumes using efficient algorithmic approaches because with increasing dataset there are chances of finding those patterns which are invalid. Solution to this problem is the use of prior knowledge to identify irrelevant variables.
2. There are some issues related to prompt change, deletion of data that can make previously discovered patterns invalid. Possible solutions are to discover methods for up-dating the patterns.
3. Problem of missing and noisy data: This problem is related to business database and mostly happens when KDD methods and tools do not easily incorporate prior knowledge about a problem.

Applications of Data Mining:

1. **Healthcare:** The past decade has seen an explosive growth in biomedical research, ranging from the development of new pharmaceuticals and in cancer therapies to the identification and study of human genome

by discovering large scale sequencing patterns and gene functions. Recent research in DNA analysis has led to the discovery of genetic causes for many diseases and disabilities as well as approaches for disease diagnosis, prevention and treatment.

2. **Telecommunication:** The telecommunication industry has quickly evolved from offering local and long distance telephone services to provide many other comprehensive communication services including voice, fax, pager, cellular phone, images, e-mail, computer and web data transmission and other data traffic. The integration of telecommunication, computer network, Internet and numerous other means of communication and computing are underway. Moreover, with the deregulation of the telecommunication industry in many countries and the development of new computer and communication technologies, the telecommunication market is rapidly expanding and highly competitive. This creates a great demand from data mining in order to help understand business involved, identify telecommunication patterns, catch fraudulent activities, make better use of resources ,and improve the quality of service.
3. **Finance:** Most banks and financial institutions offer a wide variety of banking services (such as checking, saving, and business and individual customer transactions), credit (such as business, mortgage, and automobile loans), and investment services (such as mutual funds). Some also offer insurance services and stock services. Financial data collected in the banking and financial industry is often relatively complete, reliable and high quality, which facilitates systematic data analysis and data mining. For example it can also help in fraud detection by detecting a group of people who stage accidents to collect on insurance money.
4. **Data Preprocessing:** To identify useful novel patterns in distributed, large, complex and temporal data, data mining techniques has to evolve in various stages. The present techniques and algorithms of data preprocessing stage are not up to the mark compared with its significance in finding out the novel patterns of data. In future there is a great need of data mining applications with efficient data preprocessing techniques.
5. **Complex object of data:** Data mining is going to penetrate in all fields of human life; the presently available data mining techniques are restricted to mine the traditional forms of data only, and in future there is a potentiality for data mining techniques for complex data objects like high dimensional, high speed data streams, sequence, noise in the time series, graph, Multi-instance objects, Multi-represented objects and temporal data.
6. **Scientific Computing:** In recent years data mining has attracted the research in various scientific computing applications, due to its efficient analysis of data, discovering meaningful new correlations, patterns and trends with the help of various tools and techniques. More research has to be done in mining of scientific data in particular approaches for mining astronomical, biological, chemical, and fluid dynamical data analysis. The ubiquitous use of embedded systems in sensing and actuation environments plays major impending developments in scientific computing will require a new class of techniques capable of dynamic data analysis in faulty, distributed framework. The research in data mining requires more attention in ecological and environmental information analysis to utilize our natural environment and resources. Significant data mining research has to be done in molecular biology problems.
7. **Data mining using multimedia:** The multimedia data includes images, video, audio, and animation. Data mining techniques followed in multimedia data are rule based decision tree classification algorithms like Artificial Neural Networks, Instance-based learning algorithms, Support Vector Machines, Association rule mining, clustering methods.
8. **Emergence of Data mining in other fields:** Other data mining areas include visualization, medical, pattern, wireless networks, association rule based mining.
9. **Data mining using multimedia:** The multimedia data includes images, video, audio, and animation. Data mining techniques followed in multimedia data are rule based decision tree classification algorithms like Artificial Neural Networks, Instance-based learning algorithms, Support Vector Machines, Association rule mining, clustering methods.
10. **Data mining is used for market basket analysis:** Data mining technique is used in MBA (Market Basket Analysis).When the customer want to buying some products then this technique helps us finding the associations between different items that the customer put in their shopping buckets. Here the discovery of such associations that promotes the business technique .In this way the retailers uses the data mining technique so that they can identify that which customers intension (buying the different pattern).In this way this technique is used for profits of the business and also helps to purchase the related items.
11. **Data mining is now used in many different areas in manufacturing engineering:** When we retrieve the data from manufacturing system then the customer is to use these data for different purposes like to find the errors in the data,to enhance the design methodology ,to make the good quality of the data ,how best the data can be supported for making the decision . But most of time the data can be first analyzed then after find the hidden patterns which will be control the manufacturing process which will further enhance the quality of the products .Since The importance of data mining in manufacturing has clearly increased over the last 20 years, it is now appropriate to critically review its history and Application.
12. **In language research and language:** engineering much time extra linguistic information is needed about a text. A linguistic profile that contains large number of linguistic features can be generated from text file automatically using data mining. This technique found quite effective for authorship verification and recognition. A profiling system using combination of lexical and syntactic features shows 97% accuracy in selecting correct author for the text. The linguistic profiling of text effectively used to control the quality of

language and for the automation language verification. This method verifies automatically the text is of native quality. The results show that language verification is indeed possible.

REFERENCES

[1] S. Mitra, S. K. Pal, and P. Mitra. "Data mining in soft computing framework: A survey", *IEEE Trans. Neural Networks*, vol. 13, pp. 3 - 14., 2006.
[2] Han, J., & Kamber, M. 2001. Data mining: Concepts and techniques .Morgan-Kaufman Series of Data Management Systems. San Diego: Academic Press.
[3] Z. K. Baker and V. K.Prasanna. "Efficient Parallel Data Mining with the Apriori Algorithm on FPGAs" *IEEE International Parallel and Distributed Processing Symposium* (IPDPS '05), 2005.
[4] Chris Clifton, Wei Jiang, M. Murugesan, and M.E. Nergiz, "Is Privacy Still and Issue for Data Mining", In NGDM, Taylor and Francis, 2008.
[5] J. R. Quinlan. *C4.5: Programs for Machine Learning*, San Francisco: Morgan Kaufmann Publishers, 1993.
[6] T. M. Mitchell, "Generalization as Search", in *Artificial Intelligence* vol. 18 no. 2, pp.203-226. 1982.
[7] Kotsiantis, S., Kanellopoulos, D., Pintelas, P., "Multimedia mining", *WSEAS Transactions on Systems*, No 3, s. 3263-3268, 2005.
[8] Shonali Krishnaswamy, "Towards Situationawareness and Ubiquitous Data Mining for Road Safety: Rationale and Architecture for a Compelling Application", *Proceedings of Conference on Intelligent Vehicles and Road Infrastructure 2005*, pages-16, 17. Available at: http://www.csse.monash.edu.au/~mgaber/CameraReadyI .
[9] Abdulvahit, Torun. , Ebnem, Düzgün, "Using spatial data mining techniques to reveal vulnerability of people and places due to oil transportation and accidents: A case study of Istanbul strait", *ISPRS Technical Commission II Symposium*, Vienna. Addison Wesley, 1st edition. 2006.
[10] O. Goldreich. The Foundations of Cryptography, Vol. 2, Chapter 7. Cambridge University Press, Cambridge, UK, 2004.
[11] James E. Gentle, "Challenges in Financial Data Mining", In Next Generation of Data Mining, Taylor and Francis Group, LLC 2008.
[12] Olfa Nasraoui and Maha Soliman, "Market-Based Profile Infrastructure: Giving Back to the User", Next Generation of Data Mining, Taylor and Francis, 2008.
[13] C. Potter, P.-N. Tan, V. Kumar, C. Kucharik, S. Klooster, V. Genovese, W. Cohen, and S. Healey. "Recent history of large-scale ecosystem disturbances in North America derived from the AVHRR satellite record". *Ecosystems*, 8(7):808–824, 2005.
[14] Chris Clifton, Wei Jiang, M. Murugesan, and M.E. Nergiz, "Is Privacy Still and Issue for Data Mining", In NGDM, Taylor and Francis, 2008.
[15] International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.4, No.5, September 2014.
[16] International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.4, No.5, September 2014.
[17] IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 5, No 1, September 2011.
[18] Data Mining Software at: http://www.dataminingsoftware.com.
[19] M.S. Chen, J. Han, and P.S. Yu. "Data mining: An overview from database perspective", *IEEE transactions on Knowledge and Data Eng*., 8(6):866-883, December 1999.
[20] A. Kobsa. "Technical solutions for Privacy-enhanced personalization", *Communications of the ACM*, 50: 24–33, August 2007.
[21] International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.3, June 2012.
[22] Johina et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (3) , 2015, 2928-2930
[23] International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 4, April 2015.

