An Empirical Study of Techniques and various Domains in Data mining for efficient approach in various fields

Dr. Dinesh Singh

Associate Professor, Department of Computer Applications, TERI PG College, Ghazipur

Abstract-Data mining extracts the knowledge and information from a large volume of data which may be stored in multiple heterogeneous data base. Due to tremendous grow of data, need arises to give a viewfor a research field called Data Mining, which attract attention from researchers fromvarious fields which includes Database Designers, Statisticians, Pattern Recognizer, Machine Learning, and Data Visualization. Knowledge or information is conveying the message through direct or indirect. Efficient techniques can be developed and tailored for solving complex problems using data mining in various fields including agriculture, engineering, health care, Security, sports, knowledge support systems etc. Data mining may be defined as the process of exploration and analysis of large quantities of data in order to discover meaningful patterns and rules. Manually analyzing, classifying, and summarizing the data is impossible because of the incredible increase in data in this age of net work and information sharing. Data mining is also stated as essential process where intelligent methods are applied in order to extract the data patterns. This paper discus the fundamentals of data mining and current research of data mining for developing new techniques with efficient approach in knowledge discovery system.

Keywords: Data Mining, Knowledge Discovery System, Exploration

I. INTRODUCTION

Data Mining refers to extract useful information from vast amounts of data. Many other terminologies are being used to interpret data mining, such as knowledge mining from databases, knowledge extraction, data analysis, and data archaeology. Nowadays, it is commonly agreed that data mining is an essential step in the process of knowledge discovery in databases, or KDD. In this paper, based on a broad view of data mining functionality, data mining is the process of discovering interesting knowledge from large amounts of data stored either in databases, data warehouses, or other information repositories. Web mining is the application of data mining techniques to extract useful knowledge from web data that includes web documents, hyperlinks between documents, usage logs of web sites, etc. This technique enables an individual or a company to promote business, understanding marketing dynamics, new promotions floating on the Internet, etc. The conventional and traditional system of data analysis in agriculture is purelydependent on statistics. Data mining is a modern data analysis technique. It has wide range of applications in the field of agriculture.Humans have been manually extracting patterns from data for centuries, but the increasing volume of data in modern times has called for more automated approaches. Information leads to power and success, and thanks to sophisticated technologies such as computers, satellites, etc., we have been collecting tremendous amounts of information. Initially, with the advent of computers and means for mass digital storage, we started collecting and storing all sorts of data, counting on the power of computers to help sort through this amalgam of information. Unfortunately, these massive collections of data stored on disparate structures very rapidly became overwhelming. A variety of information collected in digital form in databases and in flat files. business transactions, scientific data, medical and personal data, surveillance video and pictures, satellite sensing, games, digital media, CAD and software engineering data, virtual worlds, text reportsand e-mail messages, The World Wide Web repositories. Early methods of identifying patterns in data include Bayes' theorem and regression analysis. The proliferation, ubiquity and increasing power of computer technology has increased data collection and storage. As data sets have grown in size and complexity, direct hands-on data analysis has increasingly been augmented with indirect, automatic data processing. This has been aided by other discoveries in computer science, such as neural networks, clustering, genetic algorithms, Data mining is a field which is concerned with understanding data. In other words, the aim is to look for patterns in data. There are number of commercial data mining system available today yet there are many challenges in this field.

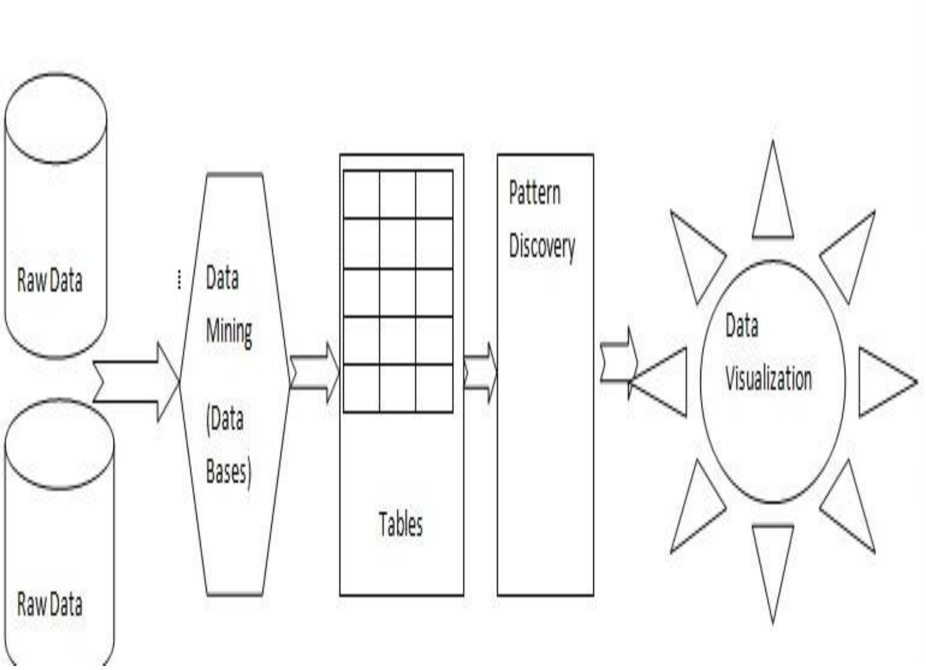


Figure 1: Data Mining Process

In step 1, extraction from data repository, step 2 involves loading of data in data bases step 3 includes Data transformation step 4 involves pattern recognition using various algorithm and step 5 includes Data interpretations and visualization of result.

II. LITRETURE REVIEW

Necessity is the requirement of invention. And from the ancient times, our ancestors have been searching for useful information from data by hand[5]. However, with the rapidly increasing in the volume of the data during 1950s, volume of data in modern times, more automatic and effective mining approaches are required. Few methods such as Bayes' theorem in the 1700s, regression analysis in the 1800s were some of the first techniques used to identify patterns in data. In thevarious decades of 1900s, with the proliferation, ubiquity, and continuously developing power of computer technology, data collection and data storage were remarkably enlarged. As data sets have grown in size and complexity, direct hands-on data analysis has increasingly been augmented with indirect, automatic data processing. This has been aided by other discoveries in computer science, such as neural networks, clusteringgenetic algorithms Decision trees in the 1960s and support vector machines.

Data mining is the process of applying these methods to data with the intention of uncovering hidden patterns. Data mining or data mining technology has been used for many years by many fields such as businesses, scientists and governments. It is used to sift through volumes of data such as airline passenger trip information, population data and marketing data to generate market research reports, although that reporting is sometimes not considered to be data mining.

Data mining commonly involves four classes of tasks[2]

(1) Classification, arranges the data into predefined groups;

(2) Clustering, is like classification but the groups are not predefined, so the algorithm will try to group similar items together;

(3) Regression, attempting to find a function which models the data with the least error;

(4) Association rule learning, searching for relationships between variables.

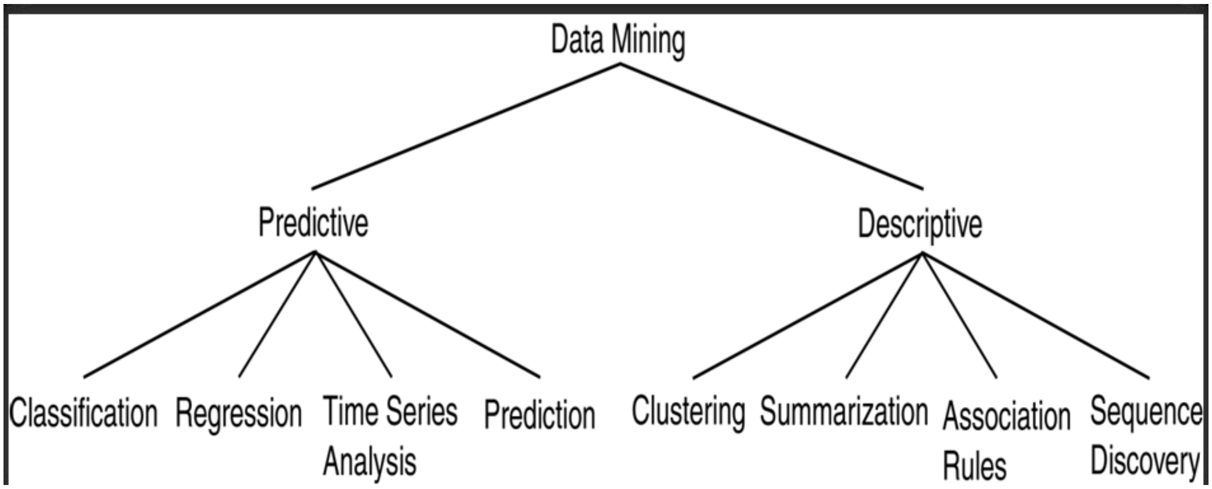
Data mining functionalities include data characterization, data discrimination, association analysis, classification, clustering, outlier analysis, and data evolution analysis. Data characterization is a summarization of the general characteristics or features of a target class of data. Data discrimination is a comparison of the general features of target class objects with the general features of objects from one or a set of contrasting classes. Association analysis is the discovery of association rules showing attribute-value conditions that occur frequently together in a given set of data. Classification is the process of finding a set of models or functions that describe and distinguish data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown. Clustering analyzes data objects without consulting a known class model. Outlier and data evolution analysis describe and model regularities or trends for objects whose behavior changes over time.

Cluster-based Temporal Mobile Sequential Pattern Mine (CTMSP-Mine), is used to discover the Cluster-based Temporal Mobile Sequential Patterns (CTMSPs)[1]. In CTMSP-Mine technique user clusters are constructed by a novel algorithm named Cluster-Object-based Smart Cluster Affinity Search Technique (CO-Smart-CAST) and similarities between users are evaluated by the proposed measure, Location-Based Service Alignment. Four important research issues namely clustering of mobile transaction sequences, Time segmentation of mobile transaction sequences.

It predict subsequent behaviors according to user's previous mobile transaction sequences and current time mining an interdisciplinary subfield of computer science is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems.

III. DATA MINING TECHNIQUES

Data mining means collecting relevant information from unstructured data. So it is able to help achieve specific objectives. The purpose of a data mining effort is normally either to create a descriptive model or a predictive model .A descriptive model presents, in concise form, the main characteristics of the data set. The purpose of a predictive model is to allow the data miner to predict an unknown value of a specific variable; the target variable.



3.1. Overview of Mining Techniques

3.1.1 Data mining

Data mining is the process of non-trivial discovery of useful knowledge from implied, previously unknown, and potentially useful information from data in large databases. Hence, it is called as a core element in knowledge discovery, often used synonymously. The data is integrated and cleaned so that the relevant data is retrieved. Data mining presents discovered data that is not just clear to data mining analysts but also for domain experts who may use it to derive actionable recommendations. Successful applications of data mining include the analysis of genetic patterns, graph mining in finance, expert system to get proper advice and consumer behavior in marketing. Traditional data mining uses structured data stored in relational tables, spreadsheets, or flat files in the tabular form. With the growth of the Web and text documents, Web mining and text mining are becoming increasingly important and popular[3].

3.1.2 Web Mining

Web mining is the term of applying data mining techniques to automatically discover and extract useful information from the World Wide Web documents and services. Web mining is the application of data mining techniques to extract knowledge from Web data, including Web documents, hyperlinks between documents, usage logs of web sites, etc. Web mining can be broadly divided into three distinct categories, according to the kinds of data to be mined that are web content mining, web structure mining and web usage mining.

3.1.3. The k-means approach

The k-means is a data mining technique for clustering. Given a set of data with unknown classification, the aim is to find a partition of the set in which similar data are grouped in the same cluster. Themeasure of similarities between data samples is provided using a suitable distance: samples that areclose to each other are considered similar. The parameter k in the k-means algorithm plays animportant role as it specifies the number of clusters in which the data must be partitioned.The idea behind the k-means algorithm is quite simple. Given a certain partition of the data in kclusters, the centers of the clusters can be computed as the mean of all samples belonging to a cluster.The center of the cluster can be considered as the representative of the cluster, because the center isquite close to all samples in the cluster, and therefore it is similar to all of them. It follows that a clustercontains similar data if all its samples are closer to its center and not to the center of some other cluster.Therefore, when samples belonging to a cluster are closer to the center of a different cluster, the kmeansalgorithm moves the corresponding data samples from their original cluster to the new cluster.

IV. CLASSIFICATION[7]

Classification is the most commonly applied data mining technique, which employs a set of pre-classifiedexamples to develop a model that can classify the population of records at large. Fraud detection and creditriskapplications are particularly well suited to this type of analysis. This approach frequently employsdecision tree or neural network-based classification algorithms. The data classification process involveslearning and classification. In Learning the training data are analyzed by classification algorithm. Inclassification test data are used to estimate the accuracy of the classification rules. If the accuracy isacceptable the rules can be applied to the new data tuples. For a fraud detection application, this wouldinclude complete records of both fraudulent and valid activities determined on a record-by-record basis.The classifier-training algorithm uses these pre-classified examples to determine the set of parameters n required for proper discrimination. The algorithm then encodes these parameters into a model called aclassifier.

4.1 Types of classification models:

Classification by decision tree induction
Bayesian Classification
Neural Networks
Support Vector Machines (SVM)
Classification Based on Associations

4.2 Clustering

Clustering can be said as identification of similar classes of objects. By using clustering techniques we canfurther identify dense and sparse regions in object space and can discover overall distribution pattern andcorrelations among data attributes. Classification approach can also be used for effective means ofdistinguishing groups or classes of object but it becomes costly so clustering can be used as preprocessingapproach for attribute subset selection and classification. For example, to form group of customers based onpurchasing patterns, to categories genes with similar functionality.

4.3 Types of clustering methods

Partitioning Methods
Hierarchical Agglomerative (divisive) methods
Density based methods
Grid-based methods
Model-based methods

4.4 Web Information Extraction

Because web data are semi-structured or even unstructured, which cannot bemanipulated by traditional database techniques, it is imperative to extract web datato port them into databases for further handling[6]. The purpose of Web InformationExtraction (IE) in our web mining research support system is to extract a specificportion of web documents useful for a research project. A specific web data set canbe scattered among different web hosts and have different formats. IE takes web documents as input, identities a core fragment, and transforms that fragment into astructured and unambiguous format. This chapter describes the design of a webIE system

4.5 Wrappers

Information extraction on the web brings us new challenges. The volume of webdocuments is enormous, documents change often, and contents of documents aredynamic. Designing a general web IE system is a hard task. Most IE systems usea wrapper to extract information from a particular web site. A wrapper consists of a set of extraction rules and specific code to apply rules to a particular site. Awrapper can be generated manually, semi-automatically or automatically.

The manual generation of a wrapper requires writing ad hoc codes based on then creator's understanding of the web documents. Programmers need to understandthe structure of a web page and write codes to translate it. Techniques have beendeveloped to use expressive grammars which describe the structure of a web page,and to generate extraction codes based on the specified grammar. the integration of heterogeneous information sources. A simple self-describing(tagged) object model is adapted to convert data objects to a common informationmodel. Languages and tools are developed to support the wrapping process.The manual way of wrapper generation cannot adapt to the dynamic changesof web sites. If new pages appear or the format of existing sources is changed, awrapper must be modified to adapt the new change.Semi-automatic wrapper generation uses heuristic tools to support wrapper generation.In such a system, sample pages are provided by users to hypothesize theunderlying structure of the whole web site. Based on the hypothesized structure,wrappers are generated to extract the information from the site. This approach doesnot require programmer knowledge about web pages, but demonstration of samplepages are required for each new site.Wrappers can be generated automatically by using machine learning and datanmining techniques to learn extraction rules or patterns. These systems can trainthemselves to learn the structure of web pages. Learning algorithms must be developedto guide the training process.

4.6 Wrapper Generation Tools

Many tools have been created to generate wrappers. Such tools include Languagesfor Wrapper Development, NLP-based Tools, Modeling-based Tool.Some languages are specifically designed to assist users to address the problemof wrapper generation.It also provides an explicit procedural mechanismfor handling exceptions inside the grammar parser. Web-OQL (Object QueryLanguage) is a declarative query language which aimed at performing SQL-likequeries over the web. A generic HTML wrapper parses a web page and producesan abstract HTML syntax tree. Using the syntax of the language, users can writequeries to locate data in the syntax tree and output data in some formats, i.e.,tables. Such tools require users to examine web documents and write a program toseparate extraction data.Natural language processing (NLP) techniques are used to learn extraction rulesexisting in natural language documents. These rules identify the relevant informationwithin a document by using syntactic and semantic constraints.

V. DATA MINING APPLICATION[9]

Various field adapted data mining technologies because of fast access of data and valuable information from a large amount of data. Data mining application area includes marketing, telecommunication, fraud detection, finance, and education sector, medical and so on. Some of the main applications listed below:

5.1 Data Mining in Education Sector:

Data mining can be applied in education sector then new emerging field called “Education Data Mining”. Using these term enhances the performance of student, drop out student, student behavior, which subject selected in the course. Data mining in higher education is a recent research field and this area of research is gaining popularity because of its potentials to educational institutes. Use student’s data to analyze their learning behavior to predict the results. [10]

5.2 Data Mining in Banking and Finance:

Data mining has been used extensively in the banking and financial markets. In the banking field, data mining is used to predict credit card fraud, to estimate risk, to analyze the trend and profitability. In the financial markets, data mining technique such as neural networks used in stock forecasting, price prediction and so on.

5.3 Data Mining in Market Basket Analysis:

These methodologies based on shopping database. The ultimate goal of market basket analysis is finding the products that customers frequently purchase together. The stores can use this information by putting these products in close proximity of each other and making them more visible and accessible for customers at the time of shopping .

5.4 Data Mining in Earthquake Prediction:

Predict the earthquake from the satellite maps. Earthquake is the sudden movement of the Earth's crust caused by the abrupt release of stress accumulated along a geologic fault in the interior. There are two basic categories of earthquake predictions: forecasts (months to years in advance) and short-term predictions (hours or days in advance).

5.5 Data Mining in Bioinformatics:

Bioinformatics generated a large amount of biological data. The importance of this new field of inquiry will grow as we continue to generate and integrate large quantities of genomic, proteomic, and other data.

5.6 Data Mining in Telecommunication:

The telecommunications field implement data mining technology because of telecommunication industry have the large amounts of data and have a very large customer, and rapidly changing and highly competitive environment. Telecommunication companies uses data mining technique to improve their marketing efforts, detection of fraud, and better management of telecommunication networks.

5.7 Data Mining in Agriculture:

Data mining than emerging in agriculture field for crop yield analysis a 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.[4]

5.8 Data Mining in 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 to perform efficient, reliable and secure services for their users.

VI. CONCLUSIONS

This paper provides a general idea of data mining, data techniques and data mining in various fields. The main objectives of data mining techniques are to discover the knowledge from active data. These applications use classification, Prediction, clustering, Association techniques and so on.

In Information Technology (IT) driven society, mining of heterogeneous data is an important issue. In this paper, a journey of research and practice from the year 1998 to 2012 is presented.This work focuses on research trends in Offline, Online and Uncertain data, useful data sources,links etc in an educational context. Different colleges/institutions affiliated to the same Universityshould adopt a single model for academic planning to strengthen the utilization of existingresources. Lastly this work can further be improved for designing Knowledge Discovery basedDecision Support System (KDDS) which will capable of giving right decision for research in Science & Technology based on the demand of the society.Once clusters are found, they can be labeled and considered as classes of models. Thus, asupervised learning technique can be used to obtain additional knowledge from these classes.

VII. FUTURE WORK

During the period that was available for this research, several ideas and concepts have emerged. However, due to time limitations only a few have been developed, implemented and tested.Nevertheless, generation of these experimental ideas are also important contributions of thisresearch. Below is a list of possible subsequent research in the areas covered by this thesis[7].They serve as starting points for future research in this field.Further work can be performed on clustering. For example, self-organizingmap (SOM) can be used instead of PCA. In addition, other clustering algorithms are available.Stability of clusters can also be studied. For example, through consensus clustering, it is possible to represent consensus of results across multiple runs of a clusteringalgorithm. It thus helps to know how the clustering algorithm (e.g. K-means) is affecting results.

VIII. REFERENCES

[1] M.Nagaraj1 , A. Venugopal, "Hybrid Temporal Sequential Pattern Mining Scheme for Mobile Services" International Journal of Computer Science and Mobile Computing A Monthly Journal of Computer Science and Information Technology ISSN 2320–088X IMPACT FACTOR: 6.017 IJCSMC, Vol. 5, Issue. 11, November 2016, pg.01 – 15

[2] Smita, Priti SharmaUse of Data Mining in Various Field: A SurveyIOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 16, Issue 3, Ver. V (May-Jun. 2014), PP 18-21

[3] Pranit B. Mohata Web Data Mining Techniques andImplementation for Handling Big Data A Monthly Journal of Computer Science and Information TechnologyISSN 2320–088XIJCSMC, Vol. 4, Issue. 4, April 2015, pg.330 – 334

[4] Dr P Jaganathan,"A study of data mining techniques to agriculture" International Journal Of Research In Information Technology, Volume 2, Issue 4, April 2014, Pg: 306- 313

[5] Yihao Li Data Mining: Concepts, Background And Methods Of Integrating Uncertainty In Data Mining

[6] Design A ProposalSubmitted to the Graduate Schoolof the University and Implementation of A Web Mining Research Support Systemof Notre Damein Partial Ful_llment of the Requirementsfor the Degree ofDoctor of PhilosophybyJin Xu, M.S.Gregory Madey Patrick Flynn, DirectorDepartment of Computer Science and EngineeringNotre Dame, Indiana

[7] Bharati M. Ramageri /"Data Mining Techniques And Applications"Indian Journal of Computer Science and EngineeringVol. 1 No. 4 301-305

[8] SandroSaitta"Data Mining Methodologies for SupportingEngineers during System Identification"THÈSE NO 4056 (2008)

[9] Smita1, Priti Sharma2,"Use of Data Mining in Various Field: A Survey Paper "IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 16, Issue 3, Ver. V (May-Jun. 2014), PP 18-21

[10] Rajni Jindal, Malaya Dutta Borah "A Survey On Educational Data Mining AndResearch Trends" "International Journal of Database Management Systems (IJDMS)"Vol.5, No.3, June 2013 DOI : 10.5121/ijdms.2013.5304 53