Recent Trends in Data Mining

Rakesh Suryawanshi[#1], Vaibhav Survase[*2]
[#]*Department Of Computer Engineering, University Of Mumbai*
A.C. Patil College of Engineering, Kharghar, Maharashtra, India
[1]rakeshsuryawanshi@gmail.com
[*]*Student at A.C. Patil College of Engineering*
Kharghar, Navi Mumbai, Maharashtra, India
[2]*vvsurvase@gmail.com*

Abstract— ***With the advent of computing technology, massive amount of data is generated and stored in gigantic devices or flowing into andout of the system in the form of data streams. Data mining is the knowledge discovery process by analyzing the large volumes of data from various perspective and summarizing it into useful information. Due to importance of extracting knowledge/information from the large data repositories, various applications like machine learning, Artificial intelligence, pattern recognition etc., are evolved. Hence this article discusses the current and future trends in the field of data mining.***

Keywords—**Data Mining, Recent Trends, Future trends, Knowledge Discovery.**

I. INTRODUCTION

Data Mining which is also referred to as Knowledge Discovery in Databases, is a new discipline of computer science aiming at automatic interpretation of huge datasets. Fayyad et al. describes KDD as "the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data." Additionally they define data mining as "a step in the KDD process consisting of applying data analysis and discovery algorithms."

As different types of data are available for data mining tasks, so data mining approaches poses many challenging research issues in data mining. In this paper we discuss the recent trends in Data Mining which includes exploration of new areas of applications and new ways of handling complex data types, algorithms scalability, constraint based mining and visualization methods, the integration of data mining with data warehousing and database systems, the standardization of mining languages and data privacy, protection and security.

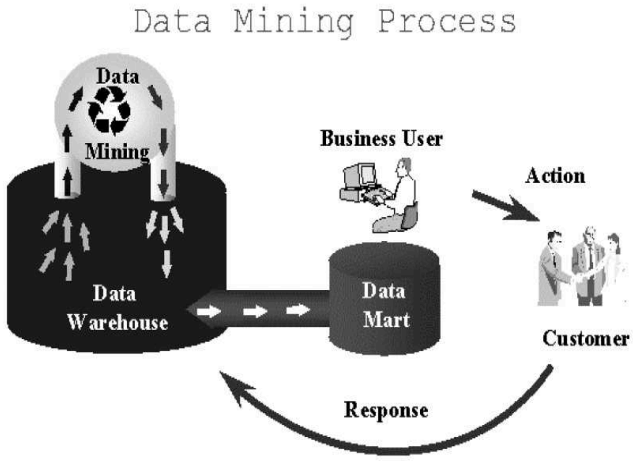
II. DATA MINING PROCESS

The data mining process can be divided into four steps. This process takes input of already summarized data found in a data warehouse and transforms into useful information. These four steps are:

1. Data Selection:
2. Data Transformation
3. Mining the Data
4. Interpretation of results

Data selection consists of gathering the data for analysis. It consists of finding and constructing pre-selection criteria before the extraction of data. It can be the most important step in the process. Data transformation will then convert the data into appropriate format. Data is further synthesized by computing certain ratios and applying algorithms to convert the data to a particular type suitable for future applied tools. Data-mining tool will extract the relevant information from the data warehouse environment. These systems need not always interact with data warehouse, if given raw data from database. The final step in the data mining process consists of interpreting the results. Once the extracted information is analyzed and interpreted, the useful results can be passed to decision maker through a DSS. If interpreted results are not satisfactory, above steps are repeated until the information generated contains maximum added value to data miner.

Figure 1: Data Mining Process



III. CURRENT TRENDS IN DATA MINING

The data mining applications is enormously growing in fields like health care, finance, security, retail,

telecommunication, risk analysis, etc. The increasing complexities, challenges and improvements in technology have led to new challenges to data mining. The various challenges includes different data formats, distributed data mining, networking and computation resources, growing business challenges, etc. Some of the current trends in data mining applications are below in TABLE 1.

Table 1: Current Data Mining areas and techniques to mine the various Data Formats

Data mining type	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
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.

IV. RECENT TRENDS IN DATA MINING

The massive growth of data is due to the availability of data in automated form from various sources like WWW, Buisness, science, society, etc. But data is useless, if it cannot deliver knowledge.

As different types of data are available for data mining tasks, so data mining approaches poses many challenging research issues in data mining. The important applications of data mining today are mostly association and sequential pattern tools, classifiers, visualization, and clustering. Collaborative filtering is new which is used in web mining. In each of these, data mining differs in the approach taken to solve problems.

A. Association and Sequencing

The industries in which association and sequencing data mining is used are health care, insurance, finance and retail industries. There are many applications in care management, procedure interactions and pharmaceutical interactions where these tools can find patterns in transaction data. In financial service sector, finding pattern in stocks can be very useful to broker optimizing the values of stocks for clients. In the telecommunications and insurance sectors, detecting fraud is a serious concern for companies in these industries. Trying to find patterns in which fraud detectors take advantage of the telephone lines is valuable. Another type of fraud common today is that of credit card fraud. Detecting when certain crimes in the credit card industry occur is important.

Since longtime association analysis is used to examine which products are frequently sold together and based on this analysis retailers can choose the optimal layouts to maximize profit. But in this area, data mining has hardly grown out of its infancy. Several retailers experiment with self-scanners. With traditional shopping, we can only find out which product is brought by a certain customer. By promoting self scanning, other important information can be obtained: how a customer wanders in their shop. This provides new possibility in data mining. We expect that real-time discounts will make their entry in the coming years. While scanning a product, a discount that is valid for certain amount of time is offered for a related product. Also attaching RFID-tags to cattle enables an efficient control of the food chain. In courier services and stock management centers this technology can lead to considerable efficiency improvements. We can also expect the advent of location based services or L-commerce. L-commerce aims at providing the relevant information with the use of data about the current location of the person.

B. Bioinformatics

Bio-informatics is the science concerning management, mining and interpretation of biological sequences and structures. Analysis of genes and proteins which are very huge in size and very sophisticated in function, is of much greater challenges than traditional data analysis methods. Progress in technologies such as microarrays, resulted in the beginning of the subdomains of genomics and proteomics. Lots of data is being generated, data that must be mined if mankind ever wants to expose the mysteries of cells.

Past few years, progress has been achieved in this field, but substancial research is needed to produce powerful mining tools in biological and bioinformatics field. With the advent of this field, data mining will help in understanding gene expression, the development of medicines and other problems in genomics and proteomics.

In molecular biology, many complex data mining tasks exist,which cannot be handled by standard data mining algorithms. These problems involve many different aspects, such as DNA, chemical properties, 3D structures, and functional properties. There is much interesting knowledge yet to be discovered, as far as the dynamic change regularities and/or their cross- interactions are concerned. In this regard, one of the challenges today is how to deal with the problem of dynamic temporal behavioral pattern identification and prediction in: (1) very large-scale systems (e.g. global climate changes and potential "bird flu" epidemics) and (2) human-centered systems (e.g. user-adapted human-computer interaction or P2P transactions). So far three important and challenging applications for data

mining have emerged: bioinformatics, CRM/personalization and security applications. However, more explorations are needed to expand these applications and extend the list of applications.

C. Visual Data Mining

What numbers cannot show, corresponding pictures often can. A significant implicit model frequently used is that of visualization. Visualization provides analysts with visual summaries of data. This technology facilitates the ability to quickly and easily change the type of information displayed and the particular visualization chosen. There are also visual data mining tools that may facilitate interactive mining based on user's judgment of intermediate data mining results. Recently, we have developed a Data Scope system that maps relational data into 2-D maps so that multidimensional relational data can be browsed in Google map's way.[1]

The analytic mining techniques does not rely on visualization. This approach majorly falls in two categories: either they use the visual data exploration systems or they use visualization to display the results of the mining algorithm. Also, interaction mechanisms for filtering, querying and selecting data are typically required for handling large data sets. Clustering or classification are used to enhance user interpretation. Many mining techniques involve different mathematical steps that require user intervention.

Tools should be developed for mapping data and knowledge into appealing and easy-to-understand visual forms and for interactive browsing. This field should be investigated to have high performance and fast response.

Major problem is how to mine non-relational data. Thus there's need of studying data mining methods that go beyond classification and clustering.

D. Stream data mining

Stream data refers to data which flows in and out of the machine. This data includes audio, video, computer network information flow, web click streams, and satellite data flow. Such data cannot be handled by traditional database systems. This is a great challenge on effective mining of stream data. Progress has been made on efficient methods for mining frequent patterns in stream data, multidimensional analysis of stream data, stream data classification and clustering, rare event detection, etc. Algorithms are made to count frequency of infinite stream data and to track most frequent k items in continuously arriving data. Many algorithms are developed to prevent the deterioration in prediction accuracy. Also real time anomaly detection in computer network analysis should be explored.

E. Information Network Analysis

Information network analysis includes social network analysis, web community discovery, terrorist network mining, computer network analysis, and network intrusion detection. In information networks each node contains a valuable, multidimensional information and other properties. Such networks are highly dynamic, evolving and interdependent. Although single link in a network could be noisy, unreliable, and sometimes misleading, valuable knowledge can be mined among large number of links in a massive information network. It can be used for predictive modeling across multiple relations. For user-guided clustering, for effective link based clustering, for distinguishing different objects with identical names, and for finding reliable facts among multiple conflicting web information providers. In link based object ranking, graph structures can be utilized. Link based classification predicts the class of each object. Entity resolution determines which references in data refer to same real world entity. We can also predict the existence of links based on attributes of the objects and other links. Some of the applications of prediction are predicting links among actors in social network, predicting friendships, etc. Information network graphs can also be treated as graphs and further data mining methods can be developed. In link mining we find sub graphs. Information networks forms huge, multidimensional heterogeneous graphs, mining noisy, approximate, and heterogeneous subgraphs based on different application for the construction of application-specific networks. Study of structural properties of networks like WWW, online social networks, biological networks etc. Many domains of interest today are best described as a network of interrelated heterogeneous objects. Future researches can be done on link mining. Also in many dynamic applications, it is important to develop incremental link mining algorithms.

F. Mining text, Web, and other unstructured data

WWW is huge collection of distributed collection of news, advertisements, consumer records, financial, education, government, e-commerce and many other services. The WWW also contains huge and dynamic collection of hyper linked information, providing a huge source of data mining. Technologies in text mining process include information extraction, topic tracking, summarization, categorization, clustering and concept linkage. Information extraction looks into sequences of text and calls pattern matching. A topic tracking system keeps user profiles, and predicts other document of interest based on document user views. Text summarization reduces the length and detail of the document while retaining its main points and

overall meaning. Web mining deals with extraction of useful patterns and implicit information from given artifacts or activities on WWW. Web data mining domains are web containing mining, web structure mining, and web usage mining. The technologies that are normally used in web content mining are natural language processing, information retrieval, and text mining. Web usage mining understands the user behavior and the web structure by analyzing the web access logs of different web sites. The technologies used in web data mining are natural language processing, information retrieval, and text mining, machine learning. This domain needs lot of research to be done. Some other researches on heterogeneous information integration, information extraction, in depth web semantics analysis, etc. are carried in this domain.

G. Distributed Data Mining and Mining Multi-Agent Data

In distributed network, distributed systems at strategic locations within the network. Challenge here is to correlate the data at various locations, and discover patterns in the global data seen. The goal is to minimize the amount of data shipped from various sites to reduce communication overhead. In distributed mining, the major challenge is to mine across multiple heterogeneous data sources. In a growing number of domains-email spam, counter-terrorism, intrusion detection, click spam, etc. – data mining systems face adversaries that manipulate data to produce false negatives. We need to develop systems that take this into account by combining data mining with game theory.

H. Software Engineering

Software programs executions potentially generates a huge amounts of data. Since they represent the executions of program logics coded by human beings they need to be mined. Data mining researches to monitor program execution status, improve system performance, isolate software bugs, detect software plagiarism, etc. Data mining in software engineering can be divided into two categories, static analysis and dynamic analysis based on the ability of a system which can collect traces beforehand for post analysis or it must react at real time to handle data. Research in data mining needs more enhancements in machine learning, data mining, pattern recognition and statistics. It is still a rich domain for data miners to research and further develop sophisticated, scalable, and real time mining methods.

V. FUTURE TRENDS

Future trends deals with the problems related to complex object which usually mean more parameters. Also data mining algorithms with few parameters and intuitiveness when adjusting parameters will gain more importance in order to reduce user interaction. Further , it might be possible to integrate the parameters into the optimization problem. In future, it might be possible to distinguish between meaningful from meaningless parameter settings, and judge the quality of the results. Future data mining should generate a large variety of well understandable patters. Due to variations in the paramertizations, the number of possible meaningful and useful patterns will dramatically increase and thus, an important aspect is managing and visualizing these patterns. Achieving user-friendliness with transparent or even reduced parameterization is a major goal. Usability is also enhanced with finding new types of patterns that are easy to interpret, even if the input data is very complex.

VI. CONCLUSION

In this article we discussed the recent trends in data mining. We started with the classic definition of the Knowledge Discovery and data mining. We found that the field is developed in past and will continue to grow further in future. First, we started with definition of knowledge discovery and data mining. Originally data mining concentrated on vectorial data while future data will be stored in much more complex data formats. We briefly discussed the process of data mining and current trends in data mining. Then we discussed recent trends in data mining which overviewed about the current and future possibilities of researches in data mining. Future data mining applications will be capable to tune themselves as far as possible, help domain experts to integrate their knowledge into data transformation and generate variety of possible patterns. Although no human being can foretell future, we believe that there are plenty of interesting challenges ahead in data mining.

ACKNOWLEDGENT

We would like to thank the authors of the reference papers for their helpful and valuable discussions. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors.

REFERENCES

[1] Jiawei Han and Jing Gao, “Research challenges for Data mining in Science and Engineering.”

[2] Hans-Peter Kriegel, Karsten M. Borgwardt, Peer Kroger, Alexey Pryakhin, Matthias Schubert, Artur Zimek, “Future trends in data mining”

[3] Walter Alberto Aldana, “Data Mining Industry: Emerging trends and New Opportunities”

[4] Venkatadri M. and Dr. Lokanatha C. Reddy, “A review on Data Mining from past to the future”.