Current Data Mining Trends, Techniques ,Applications and Issues

Ms. TASNEEM HASAN
Computer Science & Engineering Department
ITM -College of Engineering
Nagpur, India
tasneemh@itmnagpur.ac.in

Ms. SADIA PATKA
Computer Science & Engineering Department
Anjuman College of Engineering and Technology
Nagpur, India
sadiya.patka13@gmail.com

Abstract— Due to recent technological development knowledge has been playing a significant role on human activities. Data can be stored and maintained to generate useful information. Due to the increase in the data, it is important to extract knowledge/information from such large data repositories Data mining is the knowledge discovery process by analyzing the large volumes of data from various perspectives and summarizing it into useful information. It uses machine learning, statistical and visualization techniques to discovery and present knowledge in a form which is easily comprehensible to humans. Data mining have various applications and these applications have enriched the various fields of human life including business, education, medical, text mining, web mining etc. Hence the objective of this paper is to summarize important concepts, issues of data mining, to discuss various improvements in the field of data mining from past to the present and explores the future trends. This paper imparts more number of applications of the data mining and also focuses on trends in the data mining which will helpful in the further research

Keywords- *Knowledge Discovery, Data mining, Stream mining,text mining,Web mining,*

_____*********_____

I. INTRODUCTION

With the emerging applications involving massive data e.g. customer click, retail chain transaction, stock trading, emails etc leads to the generation of continuous huge bulk of information streams. To generate useful information such massive collection of data is required. Huge size of continuously flowing data has put forward a number of challenges in data analysis.

The development of Information technology has paved way to generate large amount of databases and huge data in various areas. The research in databases and information technology has given rise to approach to store and manipulate precious data for further decision making [1].

 Data mining is a process to extract the implicit information and knowledge by extracting from the mass, incomplete, noisy, fuzzy and random data with knowing the data well in advance and which is potentially useful to various fields. All most every field of human life has become data-intensive, which made the data mining as an essential component. Hence, this paper reviews the various aspects of data mining of data mining and aims to summarize trends and related applications of data mining

II. RELATED WORK

Data mining is formally defined as the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data. 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. 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 of data mining applications are.

Data mining also called Knowledge-Discovery [2] is one of the hot topics in the field of knowledge extraction from database. It combines various research areas such as databases, machine learning, artificial intelligence, statistics, automated scientific discovery, data visualization, decision science, high performance computing etc. Today, the maturity of these techniques, coupled with high-performance relational database engines and broad data integration efforts, make these technologies practical for current data warehouse environments Data mining is ready for application in the business world because it is supported by three technologies that are now sufficiently mature[3]:

- Massive data collection.
- Powerful multiprocessor computers.
- Data mining algorithms.

A. Data privacy and Security Issues

In today's fast developing world for any form of data collection and knowledge discovery security is the major constraint. In process of data mining large amounts of sensitive and private information about individuals or companies is gathered and stored. Moreover, data mining could disclose new implicit knowledge about individuals or groups that could be against privacy policies, especially if there is potential dissemination of discovered information. Another issue that arises from this concern is the appropriate use of data mining. Due to the value of data, databases of all

387

sorts of content are regularly sold, and because of the competitive advantage that can be attained from implicit knowledge discovered, some important information could be withheld, while other information could be widely distributed and used without control [5].

B. Data collection issue

The current approach is to collect as much data as possible now and process it, or try to process it, later. The concern is whether we are collecting the right data at the appropriate amount, whether we know what we want to do with it, and whether we distinguish between what data is important and what data is insignificant. Another issue related to data sources is the subject of heterogeneous databases and the focus on diverse complex data types. We are storing different types of data in a variety of repositories. It is difficult to expect a data mining system to effectively and efficiently achieve good mining results on all kinds of data and sources. Different kinds of data and sources may require distinct algorithms and methodologies. A versatile data mining tool, for all sorts of data, may not be realistic. Moreover, the proliferation of heterogeneous data sources, at structural and semantic levels, poses important challenges not only to the database community but also to the data mining community.

C. Friendly User Interface

The knowledge discovered by data mining tools is useful as long as it is interesting, and above all understandable by the user. Good data visualization eases the interpretation of data mining results, as well as helps users better understand their needs. There are many visualization ideas and proposals for effective data graphical presentation. However, there is still much research to accomplish in order to obtain good visualization tools for large datasets that could be used to display and manipulate mined knowledge. Ultimately a system must hide technological complexity from the user. To facilitate this, new software, tools, and infrastructure development is needed in the areas of grid-supported workflow management, resource identification, allocation, and scheduling, and user interfaces.

D. Dynamic data

Data streams are fast, changing, uncertain and require fast response to incorporate changes in data and reflect it in output [6].

E. Noise

Any approach applied to data should be able to deal with noise and outliers [6].

F. Low Cost

The more powerful the data mining queries, the greater the utility of the information being gleaned from the data, and the greater the pressure to increase the amount of data being collected and maintained, which increases the pressure for faster, more powerful data mining queries. This increases pressure for larger, faster systems, which are more expensive [7].

III. DATA MINING TECHNIQUES

As discussed earlier, data mining is a powerful tool for analysis and interpretation of data for the extraction of interesting knowledge and patterns that could help in Decision making. There are wide ranges of data mining techniques; some of the important techniques are discussed below:

A. Clustering

Cluster analysis or clustering is a primary data mining tool which helps user understand natural grouping of attributes. By partitioning data streams into homogenous clusters in which objects have high intra class similarity and low inter class similarity, data miners can interpret about data properties which can then be utilized to develop efficient classification systems for new data or predictive model for hidden unforeseen events [9, 10].

B. Classification

Classification is a two step process in which it initially learns from training data to form a classifier which is then used to classify unknown samples from testing data. The classifier training algorithm uses these pre-classified examples to determine the set of parameters required for proper discrimination. The algorithm then encodes these parameters into a model called a classifier.

C. Regression Analysis

Regression Analysis is one of the techniques of multi-variable analysis by which one can predict future phenomenon in form of mathematical functions under the assumption that the collection of objects has local linearity in advance. A value is used as criterion of the similarity to combine clusters. By examining value, we combine linear clusters in one by one manner. It is also called clustering by line approach. One of the key concepts in regression analysis is Criterion variable, which is an attribute that plays a role of criterion of regression analysis given by analysts.

D. Neural Network

Neural network is a set of connected input/output units and each connection has a weight present with it. During the learning phase, network learns by adjusting weights so as to be able to predict the correct class labels of the input tuples. Neural networks have the remarkable ability to derive meaning from complicated or imprecise data and can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques. These are well suited for continuous valued inputs and outputs.

E. Association

Association is one of the best known data mining technique. In association, a pattern is discovered based on a relationship between items in the same transaction. That's the reason why association technique is also known as relation technique. The association technique is used in market basket analysis to identify a set of products that customers frequently purchase together.

IV. CURRENT TRENDS IN DATA MINING

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. [8].

388

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 of data mining applications are summarized in table I below [11]:

TABLE I. CURRENT DATA MINING AREAS AND TECHNIQUES TO MINE

Data mining type	Application Domain	Data representation Formats	Data mining 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	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

V. DATA MINING APPLICATIONS AT A GLANCE

A. Data Mining in Bioinformatics and Health care

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. 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 [12].

B. Text Mining and Web Mining

Text mining is the process of searching large volumes of documents from certain keywords or key phrases. By searching literally thousands of documents various relationships between the documents can be established. Using text mining however, we can easily derive certain patterns in the comments that may help identify a commonest of customer perceptions not captured by the other survey questions. An extension of text mining is web mining. Web mining is an exciting new field that integrates data and text mining within a website. It enhances the web site with intelligent behavior, such as suggesting related links or recommending new products to the consumer. Web mining is especially exciting

because it enables tasks that were previously difficult to implement. They can be configured to monitor and gather data from a wide variety of locations and can analyze the data across one or multiple sites [12].

C. 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.

D. Customer segmentation and targeted marketing

Data mining can be used in grouping or clustering customers based on the behaviors (like payment history, etc.), which in turn helps in customer relationship management (epiphany) and performs targeted marketing. Usually it becomes useful to define similar customers in a cluster, holding on good customers, weeding out bad customers, identify likely responders for business promotions [13].

VI. CONCLUSION

In this paper we summarized the various data mining approaches, applications, issues and trends. This review would be helpful to researchers to focus on the various issues of data mining. Data mining has importance regarding finding the patterns, forecasting, and discovery of knowledge etc., in different business domains. Data mining techniques and algorithms such as classification, clustering etc., helps in finding the patterns to decide upon the future trends in businesses to grow. At present scenario data mining has wide application domain almost in every industry where the data is generated that's why data mining is considered one of the most important frontiers in database and information systems. From the last decades, data mining and knowledge discovery applications have important significance in decision making and it has become an essential component in various organizations and fields.
.

REFERENCES

List and number all bibliographical references in 9-point Times, single-spaced, at the end of your paper. When referenced in the text, enclose the citation number in square brackets, for example [1]. Where appropriate, include the name(s) of editors of referenced books. The template will number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use "Ref. [3]" or "reference [3]" except at the beginning of a sentence: "Reference [3] was the first . . ."

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the reference list. Use letters for table footnotes.

389

Unless there are six authors or more give all authors' names; do not use "et al.". Papers that have not been published, even if they have been submitted for publication, should be cited as "unpublished" [4]. Papers that have been accepted for publication should be cited as "in press" [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

[1] Kapil Wankhade, Tasneem Hasan, Ravindra Thool "A Method For Evolving Data Streams", IEEE 978-1-4673-6217, 2013

[2] Tan Pang-Ning, Steinbach, M., Vipin Kumar. Introduction to Data Mining, , 3rd Edition, Pearson Education, 2009. ,New Delhi, ISBN: 978-81-317-1472-0

[3] Bhoj Raj Sharmaa, Daljeet Kaura and Manjub "A Review on Data Mining: Its Challenges, Issues and Applications|" International Journal of Current Engineering and Technology , Vol.3, No.2 (June 2013),ISSN 2277 - 4106

[4] Meenakshi Sharma ,"Data Mining: A Literature Survey "International Journal of Emerging Research in Management &Technology, Volume-3, Issue-2, ISSN: 2278-9359 February 2014.

[5] Jing He, "Advances in Data Mining: History and Future", Third international Symposium on Information Technology Application, 978-0-7695- 3859-4, IEEE, 2009

[6] Kapil Wankhade, Tasneem Hasan, Ravindra Thool, " A Survey: Approaches for Handling Evolving Data streams", In proceedings of International Conference on Communication Systems and Network Technologies,IEEE,INDIA,2013,pp 621-625.

[7] Baazaoui Z., H., Faiz S., and Ben Ghezala, H," A Framework for Data Mining Based Multi-Agent: An Application to Spatial Data," Proceedings of World Academy of Science, Engineering and Technology, volume 5, ISSN 1307-6884, , (2005).

[8] Venkatadari M., Dr. Lokanataha C. Reddy, "A Review on Data Mining From Past to Future", International Journal of Computer Applications, pp. 19-22, vol. 15, No. 7, Feb 2011

[9] S. Guha, N. Mishra, R. Motwani, and L. O'Callaghan, "Clustering data streams," Foundations of Computer Science, Annual IEEE Symposium, 2000, pp 359-366.

[10] Ms. Tasneem Hasan "A Data Mining Approach For Handling Evolving Data Streams,"IOSR Journal,ISSN:2278-0661,pISN:2278-8727,2014

[11] S.Hameetha Begum," Data Mining Tools and Trends – An Overview "International Journal of Emerging Research in Management &Technology ISSN: 2278-9359 February 2013

[12] Dharminder Kumar, Deepak Bhardwa,"j Rise of Data Mining: Current and Future Application Areas", IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 5, No 1ISSN (Online): 1694-0814,, September 2011

[13] Vikas Gupta, Prof. Devanand, "A survey on Data Mining: Tools, Techniques, Applications, Trends and Issues," International Journal of Scientific & Engineering Research Volume 4, Issue3, 1 ISSN 2229-5518, March-2013