ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

DATA MINING: A TECHNIQUE WITH WIDENING APPLICATIONS AND HAVING SOME ISSUES

Sonal D. Tamaskar*, Prof. Anjali B. Raut
* Student of Master of Engineering in (CS & IT) HVPM's college of Engineering and Technology Amravati, India
Associate Professor and Head of the Department of (CSE) HVPM's College of Engineering and Technology Amravati, India

ABSTRACT

Data mining, the extraction of hidden predictive information from large databases, invented as a powerful new technology with great potential to help companies and organizations to focus on the most important information in their data warehouses. Data mining uses machine learning, statistical and visualization techniques to discovery and present knowledge in a form which is easily comprehensible to humans. Various popular data mining tools and techniques are available today for supporting large amount of applications. Data mining tools predict future trends and behaviors, allowing it's users to make proactive, knowledge-driven decisions. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations. With these large amount of features and applications there are some challenging issues also which are not exclusive and are not ordered in any way. This paper presents an overview of the data mining technique, Some of its vital applications and issues needs to be addressed.

KEYWORDS: Data mining, Knowledge discovery in databases (KDD), Clustering, Spatial mining, Issues.

INTRODUCTION

The advent of information technology in various fields of human life has led to the large volumes of data storage in various formats like records, documents, images, sound recordings, videos, scientific data, and many new data formats. For better decision making, the data collected from different applications require proper mechanism of extracting knowledge/information from large data repositories. Data mining, often called as Knowledge discovery in databases (KDD), aims at the discovery of useful information from large collections of data. The important reason that attracted the attention in information technology towards field of "Data mining" is due to the perception of "we are data rich but information poor". There is huge volume of data but we are hardly able to turn them into useful information and knowledge for managerial decision making in business.

To generate information it requires massive collection of data. The data can be simple numerical figures and text documents, to more complex information such as spatial data, multimedia data, and hypertext documents. For taking complete advantage of data; the data retrieval is simply not enough, it requires a tool for automatic summarization of data, extraction of the essence of information stored, and the discovery of patterns in raw data. With the enormous amount of data stored in files, databases, data warehouses and other repositories, it is increasingly important, to develop powerful tool for analysis and interpretation of such data and for the extraction of interesting knowledge that could help in decision-making. The only answer is 'Data Mining'.

The term "Data Mining" appeared around 1990 in the database community. Data mining is the analysis step of the "Knowledge Discovery in Databases (KDD)" process [1]. It also refers to the process of retrieving knowledge by discovering novel and relative patterns from the large datasets.This is like an interdisciplinary subfield of computer science,[2][3] is the computational process of discovering patterns in large data sets involving methods consisting of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use [4].

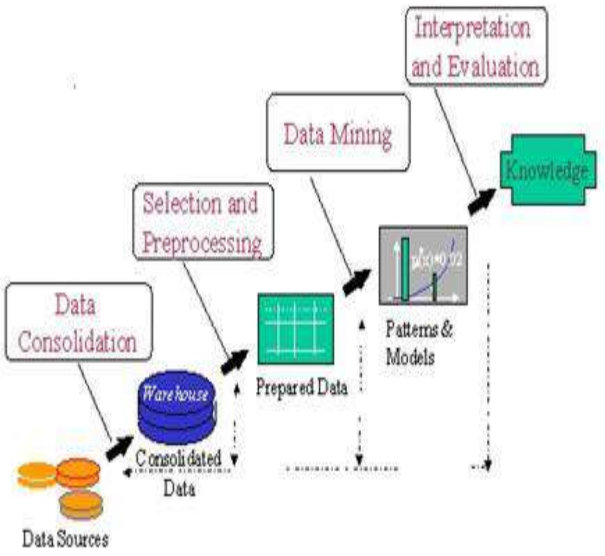

Figure 1: Data mining is the core of Knowledge discovery process.

The manual extraction of patterns from data has occurred for centuries. Early methods of identifying patterns in data include Bayes' theorem (1700s) and regression analysis (1800s). The proliferation, ubiquity and increasing power of computer technology has dramatically increased data collection, storage, and manipulation ability. As around 1990s data sets have grown in size and complexity, direct "hands-on" data analysis has increasingly been augmented with indirect, automated data processing, aided by other discoveries in computer science, such as neural networks, cluster analysis, genetic algorithms, decision trees, and support vector machines. Clustering and Classification are two distinct phases in data mining that work to provide anestablished, proven structure from a voluminous collection of facts [5]. Data mining is the process of applying these methods with the intention of uncovering hidden patterns [6] in large data sets. It bridges the gap from applied statistics and artificial intelligence that usually provide the mathematical background to database management by exploiting the way data is stored and indexed in databases to execute the actual learning and discovery algorithms more efficiently, allowing such methods to be applied to ever larger data sets. There are large no of applications some are of daily concern and some are rare but both are in same vital needs of Data mining technique. In this paper some of these applications are mention. There are also some issues related to data mining which are not exclusive and are not ordered in any way, some of these issues are also given in this paper.

TECHNOLOGICAL INFRASTRUCTURE AND WORKING OF DATA MINING

Today, data mining applications are available on all size systems for mainframe, client/server, and PC platforms. System prices range from several thousand dollars for the smallest applications up to $1 million a terabyte for the largest. Enterprise-wide applications generally range in size from 10 gigabytes to over 11 terabytes. Relational database storage and management technology is adequate for many data mining applications less than 50 gigabytes.

There are two critical technological drivers:

- Size of the database: the more data being processed and maintained, the more powerful the system required.
- Query complexity: the more complex the queries and the greater the number of queries being processed, the more powerful the system required.

While large-scale information technology has been evolving separate transaction and analytical systems, data mining provides the link between the two. Data mining software analyzes relationships and patterns in stored transaction data based on open-ended user queries. Several types of analytical software are available: statistical, machine learning, and neural networks. Generally, any of four types of relationships are sought:

- Classification: Stored data is used to locate data in predetermined groups. For example, a restaurant chain could mine customer purchase data to determine when customers visit and what they typically order. This information could be used to increase traffic by having daily specials.
- Clusters: Data items are grouped according to logical relationships or consumer preferences. For example, data can be mined to identify market segments or consumer affinities.
- Associations: Data can be mined to identify associations. The beer-diaper example is an example of associative mining.
- Sequential patterns: Data is mined to anticipate behavior patterns and trends. For example, an outdoor equipment retailer could predict the likelihood of a backpack being purchased based on a consumer's purchase of sleeping bags and hiking shoes.

More study on these relationships can be found in [7,8,9]. These relationships are then used in many techniques like: Neural networks, fuzzy logic, Intelligence agent systems, Modelling, knowledge-based systems, System optimization and Information systems, together with their applications in nearly all application domains [10] for understanding the patterns or getting more optimized solutions.

APPLICATIONS OF DATA MINING

Business and Organisations

Data mining is primarily used today by companies with a strong consumer focus - retail, financial, communication, and marketing organizations. It enables these companies to determine relationships among "internal as well as external". Internal factors such as price, product positioning, or staff skills, and external factors such as economic indicators, competition, and customer demographics. And, it enables them to determine the impact on sales, customer satisfaction, and corporate profits. Finally, it enables them to "drill down" into summary information to view detail transactional data.

With data mining, a retailer could use point-of-sale records of customer purchases to send targeted promotions based on an individual's purchase history. By mining demographic data from comment or warranty cards, the retailer could develop products and promotions to appeal to specific customer segments.

For example, WalMart is pioneering massive data mining to transform its supplier relationships. WalMart captures point-of-sale transactions from over 2,900 stores in 6 countries and continuously transmits this data to its massive 7.5 terabyte in data warehouse. WalMart allows more than 3,500 suppliers, to access data on their products and perform data analyses.

Businesses employing data mining may see a return on investment, but also they recognize that the number of predictive models can quickly become very large. Data mining can be helpful to human resources (HR) departments in identifying the characteristics of their most successful employees. Data mining is a highly effective tool in the catalog marketing industry.Catalogers have a rich database of history of their customer transactions for millions of customers dating back a number of years. Data mining tools can identify patterns among customers and help identify the most likely customers to respond to upcoming mailing campaigns.

Science and engineering

In recent years, data mining has been used widely in the areas of science and engineering, such as bioinformatics, genetics, medicine, education and electrical power engineering.

Data mining methods of biomedical data facilitated by domain ontologies,[11] mining clinical trial data,[12] and storing Electronics health records (EHR). There is adverse drug reaction surveillance, the used data mining methods to routinely screen for reporting patterns indicative of emerging drug safety issues in the WHO global database. There is 4.6 million suspected adverse drug reaction incidents[13]. Recently, similar ethodology has been developed to mine large collections of electronic health records for temporal patterns associating drug prescriptions to medical diagnoses[14].

In the area of electrical power engineering, data mining methods have been widely used for condition monitoring of high voltage electrical and power equipment. The purpose of condition monitoring is to obtain valuable information on, for example, the status of the insulation (or other important safety-related parameters) are much important. Data clustering techniques – such as the self-organizing map (SOM), they have been applied to vibration monitoring and analysis of transformer on-load tap-changers (OLTCS).

Spatial data mining

Spatial data mining is the application of data mining methods to spatial data. The end objective of spatial data mining is to find patterns in data with respect to its geography. Till the time, data mining and Geographic Information Systems (GIS) have existed as two separate technologies, each with its own methods, traditions, and approaches to visualization and for data analysis. In most contemporary GIS have only very basic spatial analysisfunctionality. The immense explosion in geographically referenced data occasioned by developments in the field of IT, digital mapping, remote sensing, and the global diffusion of GIS emphasizes the importance of developing the data-driven inductive approaches to geographical analysis and its modeling.

Sensor data mining

Wireless sensor networks can be used for facilitating the collection of data for spatial data mining for a variety of applications. For example, in air pollution monitoring [15].A characteristic of such networks is that nearby sensor nodes monitoring an environmental feature typically register similar values. This kind of data redundancy due to the spatial correlation in between sensor observations inspires the techniques for in-network data aggregation and data mining. By measuring the spatial correlation between data sampled by different sensors, a wide class of specialized algorithms have been developed to develop more efficient spatial data mining algorithms[16].

Visual data mining

In the process of turning from analog into digital, large data sets have been generated, collected, and stored discovering statistical patterns, trends and information which is hidden in data, for building the predictive patterns. Studies suggest visual data mining is faster and much more intuitive than is traditional data mining.[50][51]

Application in Surveillance

Data mining has been used by the U.S. government in the programs including the Total Information Awareness (TIA) program, in the Computer-Assisted Passenger Prescreening System (CAPPS II), Analysis, Dissemination, Visualization, Insight, Semantic Enhancement as for giving advices.[17] It is also used in Multi-state Anti-Terrorism Information Exchange (MATRIX).These programs have been discontinued because of controversy over whether they violate the 4th Amendment to the United States Constitution, although many programs that were formed under them continue to be funded by different organizations or under different names[18] In the context of combating terrorism, two particularly plausible methods of data mining mainly effectively used are "pattern mining" and "subject-based data mining."

THE ISSUES IN DATA MINING

Data mining algorithms embody techniques that have sometimes existed for many years,but have only lately been applied as reliable and scalable tools that time and again it outperform older classical statistical methods. While data mining has still in its importance, itis becoming a trend and ubiquitous. Before data mining develops into a conventional,mature and trusted discipline, many still pending issues have to be addressed [21]. Some ofthese issues which are not exclusive and are notordered in any wayare given below.

Issues related to Security:

Security is an important issue with any data collection that isshared or if it is intended to be used for strategic decision-making. In addition, when datais collected for customer profiling, the user behavior understanding, correlating personaldata with other information, etc., large amounts of sensitive and private informationabout individuals or companies is gathered and stored. This becomes controversial and giventhe confidential nature of some of this data and the potential illegal access to that private type of information. Moreover, data mining could disclose new implicit knowledge aboutindividuals or groups that could be against privacy policies, mainly if there is potentialdissemination of discovered information.

Some another important issue that arises from this concern isthe appropriate use of data mining. Due to the value of data, databases of all sorts ofcontent are regularly sold, and because of the competitive advantage that can be attainedfrom implicit knowledge discovered, some important information could be withheld,while other information could be widely distributed and used without control.

User interface issues:

The knowledge discovered by data mining tools is useful as longas it is interesting, and above all that are understandable by the user. Good data visualizationmakes it easy to interpret theresults of data mining, as well as helps users better understandtheir needs. Many data exploratory analysis tasks are significantly facilitated by their ability to see data in an appropriate visual presentation.

Now a days, there are many visualizationideas and proposals for effective data graphical presentation. However, there is still muchresearch to accomplish in order to obtain good visualization tools for large datasets thatcould be used to display and manipulate mined knowledge. Some major issues that related touser interfaces and visualization are “screen real-estate”, information rendering, andinteraction. Interactivity with the data and data mining results is vital important since it provides means for the user to focus, refine the mining tasks, as well as to picture thediscovered knowledge from different angles and also at different conceptual levels.

Issues of Mining methodology:

These issues pertain to the data mining approaches appliedand their limitations. The versatility of the mining approaches, the diversity ofdata available, the dimensionality of the domain, the broad analysis needs,the assessment of the knowledge discovered, the exploitation of background knowledgeand metadata, etc. Above are all such examples that can be as dictate mining methodology choices. There can also some different approaches may suit and solveuser’s needs differently.Most algorithms assume the data to be noise-free. This is of course a strong assumption which is required. Most datasets contain exceptions, invalid or incomplete information, etc., which maycomplicate, if not obscure, the analysis process and in many cases compromise theaccuracy of the results.

As a consequence, data preprocessing and transformation becomes vital. It is often seen as lost time, but data cleaning, as timeconsumingand frustrating as it may be, is one of the most important phases in theknowledge discovery process. Data mining techniques should be able to handle noise indata or incomplete information.More than the size of data, the size of the search space is even more decisive for datamining techniques. This is known as the curse of dimensionality. This“curse” affects so badly the performance of some data mining approaches that it isbecoming one of the most urgent issues to solve.

D.Performance issues:

Many artificial intelligence and statistical methods exist for dataanalysis and its interpretation. However, these methods were often not designed for the verylarge data sets, for with data mining is dealing today. This raisesthe issues of scalability and efficiency of the data mining methods when processingconsiderably large amount of data. Algorithms with exponential and even medium-order polynomialcomplexity cannot be of practical use for data mining. One method that is, sampling can be used for mining instead of the whole dataset.However, concerns such as completeness and choice of samples may arise.

The issue of performance also consist of incremental updating, and parallel programming. It is true that parallelism can help solve the size problem if the dataset can besubdivided and the results can be merged later. Incremental updating is important formerging results from parallel mining, or updating data mining results when new data becomes available without necessary to re-analyze the complete dataset.

Data source issues:

There are many issues related to the data sources, some are practicalsuch as the diversity of data types, while some are philosophical like the data glutproblem. We certainly have an excess of data since we already have more data than wecan handle and we are still collecting data at higher rate. If the spread of databasemanagement systems has helped increase for the gathering of information, the advent of datamining is certainly encouraging for more data harvesting. The current practice is to collect asmuch data as possible now and process it, or try to process it, later.

For the practical issues related to data sources, thereis 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 but it is difficult to expect a data mining system to effectively and efficiently achieve good mining results on all variety of data and their sources. Different kinds of data and their sources may require distinct algorithms for mining and methodologies to get proper data of which user is in need. Currently, there is a focus on relational databases and datawarehouses, but other approaches need to be pioneered for other specific complex datatypes and to handle large amount of data. A versatile data mining tool, for all sorts of data, may not be realistic. The proliferation of heterogeneous data sources, at structural and semantic levels, creates important challenges for both the database community and the data miningcommunity.

Issues in Spatial mining:

Geospatial data repositories tend to be very large for storing and also for retrieving. Moreover, existing GIS datasets are often splintered into feature and attribute components that are necessary to conventionally archived in hybrid data management systems. Algorithmic requirements differ substantially for relational data management and for topological that is releated to feature of data management[20]. Related to this is the range and diversity of geographic data formats, which present unique challenges. The digital geographic data revolution is creating new types of data formats beyond the traditional "vector" and "raster" formats. So it is incremental need for managing spatial issues.

CONCLUSION

As we have studied, Data mining is the process of discovering and distinguishing related, credential and criticalinformation from a prolific database.With the enormous amount of data stored in files, databases, and other repositories, it is increasingly important, if not necessary, to develop powerful means for analysis and perhaps interpretation of such data and for the extraction of interesting knowledge that could help in decision-making. Data Mining, also popularly known as Knowledge Discovery in Databases (KDD), refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. As we also seen some live applications where Data Mining is used as vital means.We also addressed some of the issues related to mining techniques. As form one side we are saying that data which is stored in large repositories is becoming important part of our life, so from other side it becomes much necessary for us to resolve all the issues related to Data mining and consequently of its large users.

REFERENCES

1. Fayyad, Usama; Piatetsky-Shapiro, Gregory; Smyth, Padhraic (1996). "From Data Mining to Knowledge Discovery in Databases". Retrieved 17 December 2008.

2. Y. Ramamohan, K. Vasantharao, C. KalyanaChakravarti, A.S.K.Ratnam ”A Study of Data Mining Tools in Knowledge Discovery Process” (IJSCE) ISSN: 2231-2307, July 2012
3. "Data Mining Curriculum". ACM SIGKDD. 2006-04-30. Retrieved 2011-10-28.
4. Clifton, Christopher (2010). "Encyclopædia Britannica: Definition of Data Mining". Retrieved 2010-12-09.
5. Tang, T., McCalla, G. (2005) Smart recommendation for an evolving elearning system: architecture and experiment, International Journal onE-Learning, 4(1), 105–129.,
6. ShomonaGracia Jacob and R.GeethaRamani, “EVOLVING EFFICIENT CLUSTERING AND CLASSIFICATION PATTERNS IN LYMPHOGRAPHY DATA THROUGH DATA MINING TECHNIQUES” 10.5121/ijsc3309.2012.
7. Romero, C., Ventura, S. (2007). Educational Data Mining: A Survey from 1995 to 2005. Expert Systems with Applications, 33, 125-146.
8. Mrs Anita Chaware, Educational Data Mining: An Emerging Trends in Education, review articla in International Journal of Advanced Research in Computer Science, Volume 2, No. 6, Nov-Dec 2011, ISSN No. 0976-5697
9. Shu-Hsien Liao, Pei-Hui Chu, Pei-Yuan Hsiao, “Data mining techniques and applications – A decade review from 2000 to 2011”, Expert Systems with Applications 39 (2012) 11303–11311
10. Zhu, Xingquan; Davidson, Ian (2007). Knowledge Discovery and Data Mining: Challenges and Realities. New York, NY: Hershey. pp. 163–189. ISBN 978-1-59904-252-7.
11. Zhu, Xingquan; Davidson, Ian (2007). Knowledge Discovery and Data Mining: Challenges and Realities. New York, NY: Hershey. pp. 31–48. ISBN 978-1-59904-252-7.
12. Bate, Andrew; Lindquist, Marie; Edwards, I. Ralph; Olsson, Sten; Orre, Roland; Lansner, Anders; and de Freitas, Rogelio Melhado; A Bayesian neural network method for adverse drug reaction signal generation, European Journal of Clinical Pharmacology 1998 Jun; 54(4):315–21
23.
13. Norén, G. Niklas; Bate, Andrew; Hopstadius, Johan; Star, Kristina; and Edwards, I. Ralph (2008); Temporal Pattern Discovery for Trends and Transient Effects: Its Application to Patient
14. Ma, Y.; Richards, M.; Ghanem, M.; Guo, Y.; Hassard, J. (2008). "Air Pollution Monitoring and Mining Based on Sensor Grid in London". Sensors 8 (6): 3601. doi:10.3390/s8063601.edit
15. Ma, Y.; Guo, Y.; Tian, X.; Ghanem, M. (2011). "Distributed Clustering-Based Aggregation Algorithm for Spatial Correlated Sensor Networks". IEEE Sensors Journal 11 (3): 641. doi:10.1109/JSEN.2010.2056916.edit
16. Zhao, Kaidi; and Liu, Bing; Tirpark, Thomas M.; and Weimin, Xiao; A Visual Data Mining Framework for Convenient Identification of Useful Knowledge
17. Keim, Daniel A.; Information Visualization and Visual Data Mining
18. Government Accountability Office, Data Mining: Early Attention to Privacy in Developing a Key DHS Program Could Reduce Risks, GAO-07-293 (February 2007), Washington, DC
19. "Total/Terrorism Information Awareness (TIA): Is It Truly Dead?". Electronic Frontier Foundation (official website). 2003. Retrieved 2009-03-15.
20. Healey, Richard G. (1991); Database Management Systems, in Maguire, David J.; Goodchild, Michael F.; and Rhind, David W., (eds.), Geographic Information Systems: Principles and Applications, London, GB: Longman
21. http://webdocs.cs.ualberta.ca/~zaiane/courses/cmput690/notes/Chapter1/ch1.pdf
22. Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome (2009). "The Elements of Statistical Learning: Data Mining, Inference, and Prediction". Retrieved 2012-08-07.