A Review on Data mining from Past to the Future

Venkatadri.M
Research Scholar,
Dept. of Computer Science,
Dravidian University, India.

Dr. Lokanatha C. Reddy
Professor,
Dept. of Computer Science,
Dravidian University, India.

ABSTRACT

Data and Information or Knowledge has a significant role on human activities. Data mining is the knowledge discovery process by analyzing the large volumes of data from various perspectives and summarizing it into useful information. Due to the importance of extracting knowledge/information from the large data repositories, data mining has become an essential component in various fields of human life. Advancements in Statistics, Machine Learning, Artificial Intelligence, Pattern Recognition and Computation capabilities have evolved the present day's data mining applications and these applications have enriched the various fields of human life including business, education, medical, scientific etc. Hence, this paper discusses the various improvements in the field of data mining from past to the present and explores the future trends.

Keywords

Knowledge Discovery in Databases, Data Mining, Historical Trends, Heterogeneous Data, Current Trends, Future Trends.

1. INTRODUCTION

The advent of information technology in various fields of human life has lead to the large volumes of data storage in various formats like records, documents, images, sound recordings, videos, scientific data, and many new data formats. The data collected from different applications require proper mechanism of extracting knowledge /information from large repositories for better decision making. Knowledge discovery in databases (KDD), often called data mining, aims at the discovery of useful information from large collections of data[1]. The core functionalities of data mining are applying various methods and algorithms in order to discover and extract patterns of stored data [2]. From the last two decades data mining and knowledge discovery applications have got a rich focus due to its significance in decision making and it has become an essential component in various organizations. The field of data mining have been prospered and posed into new areas of human life with various integrations and advancements in the fields of Statistics, Databases, Machine Learning, Pattern Reorganization, Artificial Intelligence and Computation capabilities etc. The various application areas of data mining are Life Sciences (LS), Customer Relationship Management (CRM), Web Applications, Manufacturing, Competitive Intelligence, Retail/Finance/Banking, Computer/Network/Security, Monitoring/Surveillance, Teaching Support, Climate modeling, Astronomy, and Behavioral Ecology etc. 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 trends of data mining and its relative areas from past to present and explores the future areas of it. This paper is organized as follows section 2 presents historical perspectives of data mining section 3 presents current trends in data mining section 4 presents future trends of data mining Section 5 presents the comparative statement of data mining trends and finally conclusion follows.

2. HISTORICAL TRENDS OF DATA MINING

The building blocks of data mining is the evolution of a field with the confluences of 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 [3]. The early day's data mining trends are as under.

2.1 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.

2.2 Computing 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.

3. 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. 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

3.1 Mining the Heterogeneous data

The following table depicts various currently employed data mining techniques and algorithms to mine the various data formats in different application areas. The various data mining areas are explained after the table1.

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.

3.1.1 Hypertext / Hypermedia data mining

The hypertext and hypermedia data is a collection of data from online catalogues, digital libraries, and online information data bases which include hyperlinks, text markups and other forms of data. Web mining is the application of data mining to discover the patterns from the Web. The important data mining technique used for hypertext and hypermedia data are Classification (supervised learning), Clustering (unsupervised learning).

3.1.2 Ubiquitous data mining

The advent of laptops, palmtops, cell phones, and wearable computer devices with increasing computational capacity and proliferation of all these devices is leading to the emergence of ubiquitous computing paradigm [4]. The Ubiquitous computing environments are subsequently giving rise to a new class of applications termed Ubiquitous Data Mining (UDM). UDM is the process of analysis of data for extracting useful knowledge from the data of ubiquitous computing [5]. Traditional data mining techniques that are drawn from the combination of ML and Statistics are presently employed in ubiquitous data mining [6].

3.1.3 Multimedia data mining

The multimedia data includes images, video, audio, and animation. The data mining techniques that are applied on multimedia data are rule based decision tree classification algorithms like Artificial Neural Networks, Instance-based learning algorithms, Support Vector Machines, also association rule mining, clustering methods [7].

3.1.4 Spatial data mining

The spatial data includes astronomical data, satellite data and space craft data. Some of the data mining techniques and data structures which are used when analyzing spatial and related types of data include the use of spatial warehouses, spatial data cubes, spatial OLAP, and spatial clustering methods [8].

3.1.5 Time series data mining

A time series is a sequence of data points, measured typically at successive times spaced at uniform time intervals. Typical examples include stock prices, currency exchange rates, the volume of product sales, biomedical measurements, weather data, etc, collected over monotonically increasing time. Rule induction algorithms such as Version Space [9], AQ15 [10], C4.5 rules [11] are presently employed in Time series data mining applications.

3.2 Utilizing the Computing and Networking Resources

Data mining has been prospered by utilizing the advanced computing and networking resources like Parallel, Distributed and Grid technologies. Parallel data mining applications have evolved using the Parallel computing, typical parallel data mining applications employ the Apriori algorithm [12]. Parallel computing and distributed data mining are both integrated in Grid technologies [13]. Grid based Support Vector Machine method is used in distributed data mining [14]. Recently, various soft computing methodologies have been applied in data mining such as fuzzy logic, rough set, neural networks, evolutionary computing (Genetic Algorithms and Genetic Programming), and support vector machines to analyze various formats of data stored in distributed databases results in a more intelligent and robust system providing a human-interpretable, low cost, approximate solution, as compared to traditional techniques [15] for systematic analysis, a robust preprocessing system, flexible information processing, data analysis and decision making.

3.3 Research and Scientific Computing Trends

The explosion in the amount data from many scientific disciplines, such as astronomy, remote sensing, bio-informatics, combinatorial chemistry, medical imaging, and experimental physics are tuning to various data mining techniques, to find out useful information. The Direct-kernel based techniques are powerful data mining tool for predictive modeling, feature selection and visualization in scientific computing [16].

3.4 Business Trends

Today's business must be more profitable, react quicker and offer high quality services that ever before. With these types of expectations and constraints, data mining becomes a fundamental technology in enabling customer's transactions more accurately. Data mining techniques of classification, regression, and cluster analysis are used for in current business trends [17]. Most of the current business data mining applications utilize the classification and prediction techniques for supporting business decisions. In business environment data mining has evolved to Decision Support Systems (DSS) and very recently it has grown to Business Intelligence (BI) systems.

4. FUTURE TRENDS

Due to the enormous success of various application areas of data mining, the field of data mining has been establishing itself as the major discipline of computer science and has shown interest potential for the future developments. Ever increasing technology and future application areas are always poses new challenges and opportunities for data mining, the typical future trends of data mining includes

- Standardization of data mining languages
- Data preprocessing
- Complex objects of data
- Computing resources
- Web mining
- Scientific Computing
- Business data

4.1 Standardization of data mining languages

There are various data mining tools with different syntaxes, hence it is to be standardized for making convenient of the users. Data mining applications has to concentrate more in standardization of interaction languages and flexible user interactions.

4.2 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.

4.3 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.

4.4 Computing Resources

The contemporary developments in high speed connectivity, parallel, distributed, grid and cloud computing has posed new challenges for data mining. The high speed internet connectivity has posed a great demand for novel and efficient data mining techniques to analyze the massive data which is captured of IP packets at high link speeds in order to detect the Denial of Service (DoS) and other types of attacks. Distributed data mining applications demand new alternatives in different fields, such as discovery of universal strategy to configure a distributed data mining, data placement at different locations, scheduling, resource management, and transactional systems etc. New data mining techniques and tools are needed to facilitate seamless integration of various resources in grid based environment. Moreover, grid based data mining has to focus seriously to address the data privacy, security and governance. Cloud computing is a great area to be focused by data mining, as the Cloud computing is penetrating more and more in all ranges of business and scientific computing. Data mining techniques and applications are very much needed in cloud computing paradigm.

4.5 Web mining

The development of World Wide Web and its usage grows, it will continue to generate ever more content, structure, and usage data and the value of Web mining will keep increasing. Research needs to be done in developing the right set of Web metrics, and their measurement procedures, extracting process models from usage data, understanding how different parts of the process model impact various Web metrics of interest, how the process models change in response to various changes that are made-changing stimuli to the user, developing Web mining techniques to improve various other aspects of Web services, techniques to recognize known frauds and intrusion detection.

4.6 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.

4.7 Business Trends

Business data mining needs more enhancement in the design of data mining techniques to gain significant advantages in today's competitive global market place (E-Business). The Data mining techniques hold great promises for developing new sets of tools that can be used to provide more privacy for a common man, increasing customer satisfaction, providing best, safe and useful products at reasonable and economical prices, in today's E-Business environment.

5. COMPARATIVE STATEMENT

The following table presents the comparative statement of various data mining trends from past to the future.

Table 2: Data mining Trends Comparative Statement

Data mining trends	Algorithms/ Techniques employed	Data formats	Computing Resources	Prime areas of applications
Past	Statistical, Machine Learning Techniques	Numerical data and structured data stored in traditional databases	Evolution of 4G PL and various related techniques	Business
Current	Statistical, Machine Learning, Artificial Intelligence, Pattern Reorganization Techniques	Heterogeneous data formats includes structured, semi-structured and unstructured data	High speed networks, High end storage devices and Parallel, Distributed computing etc…	Business, Web, Medical diagnosis etc…

Future	Soft Computing techniques like Fuzzy logic, Neural Networks and Genetic Programming	Complex data objects includes high dimensional, high speed data streams, sequence, noise in the time series, graph, Multi-instance objects, Multi-represented objects and temporal data etc…	Multi-agent technologies and Cloud Computing	Business, Web, Medical diagnosis, Scientific and Research analysis fields (bio, remote sensing etc…), Social networking etc…

6. CONCLUSION

In this paper we briefly reviewed the various data mining trends from its inception to the future. This review would be helpful to researchers to focus on the various issues of data mining. In future course, we will review the various classification algorithms and significance of evolutionary computing (genetic programming) approach in designing of efficient classification algorithms for data mining.

7. REFERENCES

[1] Heikki, Mannila. 1996. Data mining: machine learning, statistics, and databases, IEEE

[2] Fayadd, U., Piatesky -Shapiro, G., and Smyth, P. 1996. From Data Mining To Knowledge Discovery in Databases, AAAI Press / The MIT Press, Massachusetts Institute Of Technology. ISBN 0–262 56097–6 Fayap.

[3] Piatetsky-Shapiro, Gregory. 2000. The Data-Mining Industry Coming of Age. IEEE Intelligent Systems.

[4] Salmin, Sultana et al. 2009. Ubiquitous Secretary: A Ubiquitous Computing Application Based on Web Services Architecture , International Journal of Multimedia and Ubiquitous Engineering Vol. 4, No. 4, October, 2009

[5] Hsu, J. 2002. Data Mining Trends and Developments: The Key Data Mining Technologies and Applications for the 21st Century, The Proceedings of the 19th Annual Conference for Information Systems Educators (ISECON 2002), ISSN: 1542-7382. Available Online: http://colton.byuh.edu/isecon/2002/224b/Hsu.pdf

[6] Shonali Krishnaswamy. 2005. Towards Situation-awareness and Ubiquitous Data Mining for Road Safety: Rationale and Architecture for a Compelling Application (2005), Proceedings of Conference on Intelligent Vehicles and Road Infrastructure 2005, pages-16, 17. Available at : http://www.csse.monash.edu.au/~mgaber/CameraReadyI

[7] Kotsiantis, S., Kanellopoulos, D., Pintelas, P. 2004. Multimedia mining. WSEAS Transactions on Systems, No 3, s. 3263-3268.

[8] Abdulvahit, Torun. , Ebnem, Düzgün. 2006. 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.

[9] T. M. Mitchell. 1982. Generalization as Search, Artificial Intelligence, 18(2), 1982, pp.203-226.

[10] R. Michalski., I. Mozetic., J. Hong., and N. Lavrac. 1986. The AQ15 Inductive Leaning System: An Overview and Experiments, Reports of Machine Leaning and Inference Laboratory, MLI-86-6, George Maseon University.

[11] J. R. Quinlan.1992.Programs for Machine Learning, Morgan Kaufmann.

[12] Z. K. Baker and V. K.Prasanna. 2005. Efficient Parallel Data Mining with the Apriori Algorithm on FPGAs. In Submitted to the IEEE International Parallel and Distributed Processing Symposium (IPDPS '05).

[13] Jing He.2009. Advances in Data Mining: History and Future, Third international Symposium on Information Technology Application, 978-0-7695-3859-4/09 IEEE 2009 DOI 10.1109/IITA.2009.204

[14] Ali Meligy.2009. A Grid-Based Distributed SVM Data Mining Algorithm, European Journal of Scientific Research ISSN 1450-216X Vol.27 No.3. Pp.313-321 © Euro Journals Publishing, Inc. Available at : http://www.eurojournals.com/ejsr.htm

[15] S. Mitra, S. K. Pal, and P. Mitra. 2001. Data mining in soft computing framework: A survey, IEEE Trans. Neural Networks, vol. 13, pp. 3 - 14.

[16] Mark, J., Embrechts. 2005. Introduction to Scientific Data Mining: Direct Kernel Methods & Applications , Computationally Intelligent Hybrid Systems: The Fusion of Soft Computing and Hard Computing , Wiley , New York, pp. 317-365

[17] Han, J., & Kamber, M. 2001. Data mining: Concepts and techniques .Morgan-Kaufman Series of Data Management Systems. San Diego: Academic Press.

AUTHORS PROFILE

Venkatadri. M received his Masters Degree in Computer Science and Engineering from Acharya Nagarjuna University, India. He is presently pursuing his Ph.D in Computer Science, Dravidian University, India. His area of interest includes databases, data warehousing and mining, and artificial intelligence.

Dr. Lokanatha C. Reddy earned M.Sc. (Maths) from Indian Institute of Technology, New Delhi; M.Tech (CS) with Honours from Indian Statistical Institute, Kolkata; and Ph.D (CS) from Sri Krishnadevaraya University, Anantapur. Earlier worked at KSRM College of Engineering, Kadapa; Indian Space Research Organization (ISAC) at Bangalore and as the Head of the Computer Centre at the Sri Krishnadevaraya University, India; Presently, he is the Professor of Computer Science at the Dravidian University, India. His active research interests include Real time Computation, Distributed Computation, Device Drivers, Geometric Designs and Shapes, Digital Image Processing, Pattern Recognition and Networks.