ISBN 978-93-5156-328-0
International Conference of Advance Research and Innovation (ICARI-2014)

Data Mining: Pattern and Trends by Using Biocomputers

Asheesh Kumar
Department of Computer Applications, Shri Venkateshwara University, Gajraula, Uttar Pradesh, India

Article Info

Article history:
Received 2 January 2014
Received in revised form
10 January 2014
Accepted 20 January 2014
Available online 1 February 2014

Keywords

Knowledge Discovery in Databases,
Data Mining,
Historica Trends,
Heterogeneous Data,
Current Trends,
Future Trends

Abstract

This paper gives a good overview of 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.

1. Introduction

The past couple of decades have seen a dramatic increase in the amount of information or data being stored in electronic format. This accumulation of data has taken place at an explosive rate. It has been estimated that the amount of information in the world doubles every 20 months and the sizes as well as number of databases are increasing even faster. There are many examples that can be cited. Point of sale data in retail, policy and claim data in insurance, medical history data in healthcare, financial data in banking and securities, are some instances of the types of data that is being collected.

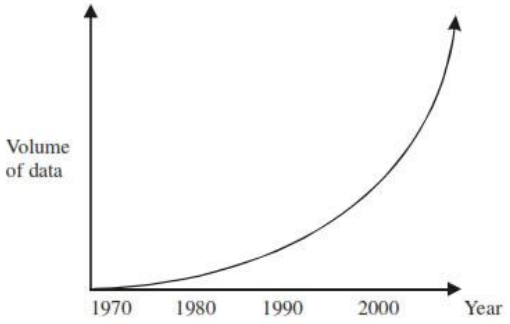

Fig: 1 The Increasing bases of Data

Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

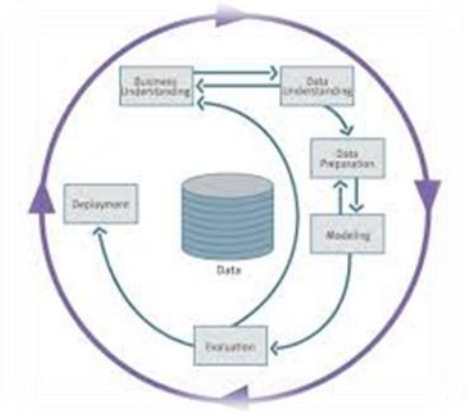

Fig: 2. A process of Data Mining

Although data mining is a relatively new term, the technology is not. Companies have used powerful computers to sift through volumes of supermarket scanner data and analyze market research reports for years. However, continuous innovations in computer processing power, disk storage, and statistical software are dramatically increasing the accuracy of analysis while driving down the cost.

Corresponding Author,
E-mail address: Choudhary.asheesh@gmail.com

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. 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.

2. Data Mining Tools

The best of the best commercial database packages are now available for data mining and warehousing including IBM DB2, INFORMIX-On Line XPS, ORACLE9i, Clementine, Intelligent Miner, 4 Thought and SYBASE System 10.

2.1 Oracle Data Mining (ODM)

Oracle enables data mining inside the database for performance and scalability. Some of the capabilities are:

An API that provides programmatic control and application integration

Analytical capabilities with OLAP and statistical functions in the database

Multiple Algorithms: Naïve Bayes and Association Rules

Real-time and Batch Scoring modes

Multiple Prediction types

Association insights

ORACLE9i Data Mining provides a Java API to exploit the data mining functionality that is embedded within the ORACLE9i database. By delivering complete programmatic control of the data base in data mining

Oracle Data Mining (ODM) delivers powerful, scalable modeling and real-time scoring. This enables e-businesses to incorporate predictions and classifications in all processes and decision points throughout the business cycle.ODM is designed to meet the challenges of vast amounts of data, delivering accurate insights completely integrated into e-business applications. This integrated intelligence enables the automation and decision speed that e-businesses require in order to compete today.

3. What can data mining do?

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" 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, Blockbuster Entertainment mines its video rental history database to recommend rentals to individual customers. American Express can suggest products to its cardholders based on analysis of their monthly expenditures.

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 Teradata data warehouse. WalMart allows more than 3,500 suppliers, to access data on their products and perform data analyses. These suppliers use this data to identify customer buying patterns at the store display level. They use this information to manage local store inventory and identify new merchandising opportunities. In 1995, WalMart computers processed over 1 million complex data queries.

3.1 How does data mining work?

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:

- **Classes**: 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.

3.2 Technology & Infrastructure

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. NCR has the capacity to deliver applications exceeding 100 terabytes. 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.

Relational database storage and management technology is adequate for many data mining applications less than 50 gigabytes. However, this infrastructure needs to be significantly enhanced to support larger applications. Some vendors have added extensive indexing capabilities to improve query performance. Others use new hardware architectures such as Massively Parallel Processors (MPP) to achieve order-of-magnitude improvements in query time. For example, MPP systems from NCR link hundreds of high-speed Pentium processors to achieve performance levels exceeding those of the largest supercomputers.

4. Issues Presented

One of the key issues raised by data mining technology is not a business or technological one, but a social one. It is the issue of individual privacy. Data mining makes it possible to analyze routine business transactions and glean a significant amount of information about individuals buying habits and preferences.

Another issue is that of data integrity. Clearly, data analysis can only be as good as the data that is being analyzed.

A key implementation challenge is integrating conflicting or redundant data from different sources. For example, a bank may maintain credit cards accounts on several different databases.

The addresses (or even the names) of a single cardholder may be different in each. Software must translate data from one system to another and select the address most recently entered.A hotly debated technical issue is whether it is better to set up a relational database structure or a multidimensional one.

In a relational structure, data is stored in tables, permitting ad hoc queries. In a multidimensional structure, on the other hand, sets of cubes are arranged in arrays, with subsets created according to category. While multidimensional structures facilitate multidimensional data mining, relational structures thus far have performed better in client/server environments. And, with the explosion of the Internet, the world is becoming one big client/server environment. Finally, there is the issue of cost. While system hardware costs have dropped dramatically within the past five years, data mining and data warehousing tend to be self-reinforcing. 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.

5. Current Trends in Data Mining

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

6. Bio computer

Biocomputers use systems of biologically derived molecules, such as DNA and proteins; to perform computational calculations involving storing, retrieving, and processing data. The development of biocomputers has been made possible by the expanding new science of nano biotechnology. The term nano biotechnology can be defined in multiple ways; in a more general sense, nano biotechnology can be defined as any type of technology that uses both nano-scale materials, i.e. materials having characteristic dimensions of 1-100 nanometers, as well as biologically based materials (34).4 A more restrictive definition views nano biotechnology more specifically as the design and engineering of proteins that can then be assembled into larger, functional structures (116-117). The implementation of nano biotechnology, as defined in this narrower sense, provides scientists with the ability to engineer biomolecular systems specifically so that they interact in a fashion that can ultimately result in the computational functionality of a computer.

Biocomputers use biologically derived materials to perform computational functions. A biocomputer consists of a pathway or series of metabolic pathways involving biological materials that are engineered to behave in a certain manner based upon the conditions (input) of the system. The resulting pathway of reactions that takes place constitutes an output, which is based on the engineering design of the biocomputer and can be interpreted as a form of computational analysis. Three distinguishable types of biocomputers include biochemical computers, biomechanical computers, and bioelectronic computers

7. Engineering bio computers

The behavior of biologically derived computational systems such as these relies on the particular molecules that make up the system, which are primarily proteins but may also include DNA molecules. Nanobiotechnology provides the means to synthesize the multiple chemical components necessary to create such a system. The chemical nature of a protein is dictated by its sequence of amino acids—the chemical building blocks of proteins.

This sequence is in turn dictated by a specific sequence of DNA nucleotides—the building blocks of DNA molecules. Proteins are manufactured in biological systems through the translation of nucleotide sequences by biological molecules called ribosomes, which assemble individual amino acids into polypeptides that form functional proteins based on the nucleotide sequence that the ribosome interprets. What this ultimately means is that one can engineer a biocomputer, i.e. the chemical components necessary to serve as a biological system capable of performing computations, by engineering DNA nucleotide sequences to encode for the necessary protein components. Also, the synthetically designed DNA molecules themselves may function in a particular biocomputer system.

Thus, implementing nanobiotechnology to design and produce synthetically designed proteins, as well as the design and synthesis of artificial DNA molecules, can allow the construction of functional biocomputers, e.g., Computational Genes.

8. Conclusion

In this paper, I have summarized about the Data mining and its trends used in Biocomputing. 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.

References

[1] 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, ages-16, 17. Available at : http://www.csse.monash.edu.au/~mgaber/Camera Ready

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

[3] Ali Meligy, A Grid-Based Distributed SVM Data Mining Algorithm, European Journal of Scientific Research, 2009 Pp.313-321 © Euro Journals Publishing, Inc http://www.eurojournals.com/ejsr.htm

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

[5] Cipolla, Emil T. Data Mining: Techniques to Gain Insight Into Your Data Enterprise Systems Journal (December 1995):18-24, 64

[6] Krivda, Cheryl D. Laps Around Business Intelligence MIDRANGE Systems, 1995

[7] Bouckaert, Remco R.; Frank, Eibe; Hall, Mark A.; Holmes, Geoffrey; Pfahringer, Bernhard;

Reutemann, Peter; Witten, Ian H. (2010). "WEKA Experiences with a Java open-source project". Journal of Machine Learning Research **11**: 2533–2541, The original title, "Practical machine learning", was changed ... The term "data mining" was [added] primarily for marketing reasons.

[8] O'Brien, J. A., & Marakas, G. M., Management Information Systems. New York, NY: McGraw-Hill/Irwin, 2011

[9] Alexander, D. (n.d.). Data Mining. Retrieved from The University of Texas at Austin: College of Liberal Arts:http://www.laits.utexas.edu/~anorman/BUS.FOR/course.mat/Alex/

[10] Goss, S. Data-mining and our personal privacy. Retrieved from The Telegraph, 2013 http://www.macon.com/2013/04/10/2429775/data-mining-and-our-personal-privacy.html

[11] Cannataro, Mario; Talia, Domenico, 2003, The Knowledge Grid: An Architecture for Distributed Knowledge Discovery". Communications of the ACM 46(1): 89–93. doi:10.1145/602421.602425. 2011

[12] Talia, Domenico; Trunfio, Paolo, How distributed data mining tasks can thrive as knowledge services". Communications of the ACM 53 (7): 132–137.doi:10.1145/1785414.1785451, 2011

[13] Seltzer, William. The Promise and Pitfalls of Data Mining: Ethical Issues

[14] Pitts, Chip, The End of Illegal Domestic Spying? Don't Count on It". Washington Spectator, 2007

[15] Taipale, Kim A., Data Mining and Domestic Security: Connecting the Dots to Make Sense of Data". Columbia Science and Technology Law Review 5 (2). OCLC 45263753. SSRN 546782, 2003

[16] Resig, John, Teredesai, Ankur, A Framework for Mining Instant Messaging Services, Proceedings of the 2004 SIAM DM Conference, 2004

[17] a b c Think Before You Dig: Privacy Implications of Data Mining & Aggregation, NASCIO Research Brief, 2004

[18] Biotech Business Week Editors, 2008 Biomedicine; Hipaa Privacy Rule Impedes Biomedical Research, Biotech Business Week, retrieved, 2009 from LexisNexis Academic

[19] 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 Records. Proceedings of the Fourteenth International Conference on Knowledge Discovery and Data Mining (SIGKDD 2008), Las Vegas, NV, pp. 963–971.

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