IJARI

International Journal of Advance Research and Innovation
Vol. 1(2), Apr-Jun 2013, pp. 20-24
Doi: 10.51976/ijari.121305

Article Info
Received: 01 Aug 2012 | Revised Submission: 22 Aug 2013 | Accepted: 28 Aug 2013 | Available Online: 20 Sept 2013

Data Mining: Review, Drifts and Issues

*Lokendra Singh**

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.

Keywords: *Knowledge Discovery in Databases; Data Mining; Historical Trends; Heterogeneous Data; Current Trends; Future Trends.*

1.0 Introduction

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.

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.

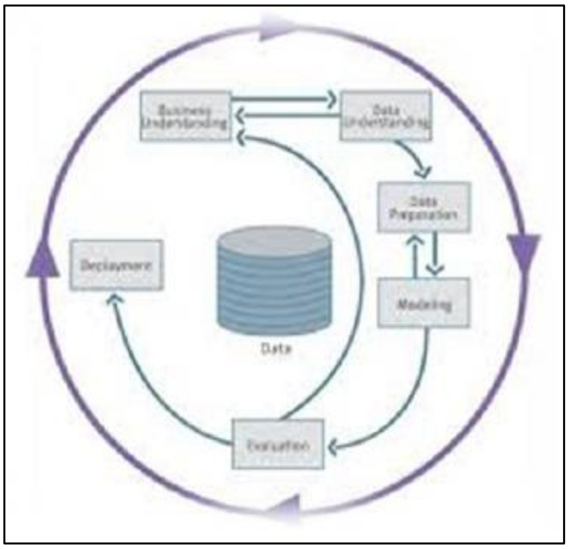


Fig 1: A Process of Data Mining

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.

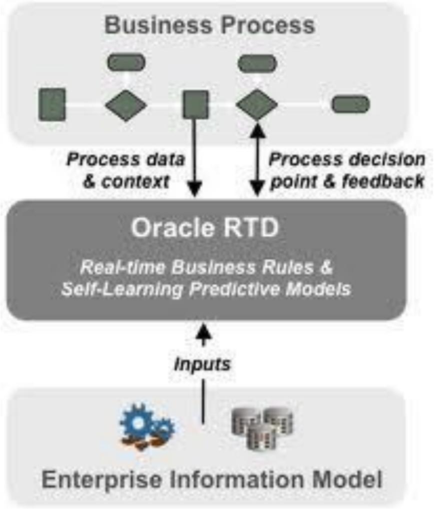
**Corresponding Author: Department of Computer Applications, Shri Venkateshwara University, Gajraula, J.P. Nagar, U. P, India (E-mail: er.anusharma18@gmail.com)*

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.

1.1. Oracle: en example

One Midwest grocery chain used the data mining capacity of Oracle software to analyze local buying patterns. They discovered that when men bought diapers on Thursdays and Saturdays, they also tended to buy beer. Further analysis showed that these shoppers typically did their weekly grocery shopping on Saturdays. On Thursdays, however, they only bought a few items.

Fig 2: Oracle: an Example of Data Mining


Business Process
Process data & context
Process decision point & feedback
Oracle RTD
Real-time Business Rules & Self-Learning Predictive Models
Inputs
Enterprise Information Model

The retailer concluded that they purchased the beer to have it available for the upcoming weekend. The grocery chain could use this newly discovered information in various ways to increase revenue. For example, they could move the beer display closer to the diaper display. And, they could make sure beer and diapers were sold at full price on Thursdays.

1.2. 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 Metadata 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.

1.3. 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.

2.0 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.

3.0 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.

4.0 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

5.0 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

1. Standardization of data mining languages
2. Data preprocessing
3. Complex objects of data
4. Computing resources 5. Web mining 6. Scientific Computing

6.0 Conclusion

References

[1] Shonali Krishna swamy 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/CameraReady

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

[3] Ali Meligy.2009. A Grid-Based Distributed SVM Data Mining Algorithm, European Journal of Scientific Research ISSN 1450-216X Vol.27 (3) Pp.313-321 © Euro Journals Publishing, Inc Available at: http://www.eurojournals.com/ejsr.htm

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

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

[6] Krivda, Cheryl D.Laps around Business IntelligenceMIDRANGE Systems (1995):32-34.

[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. (2011). Management Information Systems New York, NY: McGraw-Hill/Irwin.

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

[11] Cannataro, Mario; Talia, Domenico (2003). "The Knowledge Grid: Architecture for Distributed Knowledge Discovery". Communications of the ACM 46(1): 89–93. Doi: 10.1145 /602421. 602425 Retrieved 2011

[12] Talia, Domenico; Trunfio, Paolo (2010). "How distributed data mining tasks can thrive as knowledge services". Communications of the ACM 53 (7): 132–137.doi:10.1145/1785414. 1785451. Retrieved 2011

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

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

[15] Taipale, Kim A. (2003). "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

[16] Resig, John; and Teredesai, Ankur (2004). "A Framework for Mining Instant Messaging Services" Proceedings of the 2004 SIAM DM Conference

[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: It's 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 (3), s. 3263-3268.