**IJITKM** *Dec. 2012, Volume 6 •No. 1 • pp.—112-115 (ISSN 0973-4414)*

# DATA MINING: FOUNDATION & TECHNIQUES

**Anish Soni**

Department of CSE, HCTM, Kaithal, Haryana, India-136027

Soni_anish@yahoo.com

***Abstract*- The computing technology has significantly influenced our lives and the major impacts of this effect are Business data Processing and Scientific Computing. Earlier the development of computer techniques for business, computer professionals were concerned with designing files to store the data so that information could be efficiently retrieved. Storage sizes were restricted for storing data and also speed of accessing the data. Then the era when Database Management System simplified the task. The responsibility of various tasks, like declarative aspects of the programs was passed on to the DBA and the user could pass his query in simpler query languages. Thus almost every business began using computers for day-today activities. Then came into existence the concept of data mining on which this chapter is going put some light.**

***Keywords*—data mining, multiprocessor computers, clusters, classes, association rules.**

## Introduction

Data Mining is one of the fastest growing computer science fields. Its popularity is because of increased demand for tools that help with the analysis and understanding of huge amounts of data. Such data are generated on a daily basis by institutions like banks, insurance companies, retail stores, and on the Internet. Now a days when world is having rich sources of data need is a clear and simple methodology for extracting the knowledge that is hidden in the data.

## The Foundations of Data Mining

Data mining techniques are the result of a long process of research and product development. This process of evolution began when business data was first stored on computers, continued with improvements in data access. Data mining is ready for application in the business community because it is supported by three technologies that are now sufficiently mature:

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

Commercial databases are growing rapidly. According to a recent META Group survey of data warehouse projects it is found that 19% of respondents are beyond the 50 gigabyte level, while 59% expect to be there by second quarter of 1996. In some industries, such as retail, these numbers can be much larger. The accompanying need for improved computational engines can now be met in a cost-effective manner with parallel multiprocessor computer technology. Data mining algorithms embody techniques that have existed for at least 10 years, but have only recently been implemented as mature, reliable, understandable tools that consistently outperform older statistical methods. In the evolution from business data to business information, each new step has built upon the previous one. For example, dynamic data access is critical for drill-through in data navigation applications and the ability to store large databases is critical to data mining. From the user's point of view, the four steps listed in Table 1 were revolutionary because they allowed new business questions to be answered accurately and quickly.

The core components of data mining technology have been under development for decades, in research areas such as statistics, artificial intelligence, and machine learning. 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.

## Mechanism of Data Mining

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:

| ***Evolutionary Step*** | ***Business Question*** | ***Enabling Technologies*** | ***Product Providers*** | ***Characteristics*** |
|---|---|---|---|---|
| Data Collection (1960s) | "What was my total revenue in the last five years?" | Computers, tapes, disks | IBM, CDC | Retrospective, static data delivery |
| Data Access (1980s) | "What were unit sales in New England last March?" | Relational databases (RDBMS), Structured Query Language (SQL), ODBC | Oracle, Sybase, Informix, IBM, Microsoft | Retrospective, dynamic data delivery at record level |
| Data Warehousing & Decision Support (1990s) | "What were unit sales in New England last March? Drill down to Boston." | On-line analytic processing (OLAP), multidimensional databases, data warehouses | Pilot, Comshare, Arbor, Cognos, Microstrategy | Retrospective, dynamic data delivery at multiple levels |
| Data Mining (Emerging Today) | "What's likely to happen to Boston unit sales next month? Why? | Advanced algorithms, multiprocessor computers, massive databases | Pilot, Lockheed, IBM, SGI, numerous startups (nascent industry) | Prospective, proactive information delivery |

**Table 1. Steps in the Evolution of data mining**

- *Classes:* Stored data is used to locate data in predetermined groups. For example, a hotel 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*: Grouping of data items according to logical relationships or consumer preferences. For example, data can be mined to identify market segments or consumer affinities.

- *Associations:* Mining data 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.

## Various Techniques of Data Mining

- *Artificial neural networks:* Non-linear predictive models that learn through training and resemble biological neural networks in structure.
- *Genetic algorithms:* Optimization techniques that use process such as genetic combination, mutation, and natural selection in a design based on the concepts of natural evolution.
- *Decision trees:* Shaped structures that represent sets of decisions. These decisions generate

rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID).

- *Nearest neighbor method:* A technique that classifies each record in a dataset based on a combination of the classes of the *k* record(s) most similar to it in a historical dataset (where *k* 1). Sometimes called the *k*-nearest neighbor technique.
- *Rule induction:* The extraction of useful if-then rules from data based on statistical significance.
- *Data visualization:* The visual interpretation of complex relationships in multidimensional data. Graphics tools are used to illustrate data relationships.

## Architecture for Data Mining

For application of advanced techniques, they need to be fully integrated with a data warehouse as well as flexible interactive business analysis tools. Lots of data mining tools operate outside of the warehouse and requiring extra steps to extract, import, and analysis of data. Further, when these techniques are implemented operationally, their integration with the warehouse just simplifies the application of results from data mining. Resultant warehouse can be applied to improve business processes throughout the organization, in areas like promotional campaign management, fraud detection, new product rollout etc. Figure 1 illustrates architecture for advanced analysis in a large data warehouse.

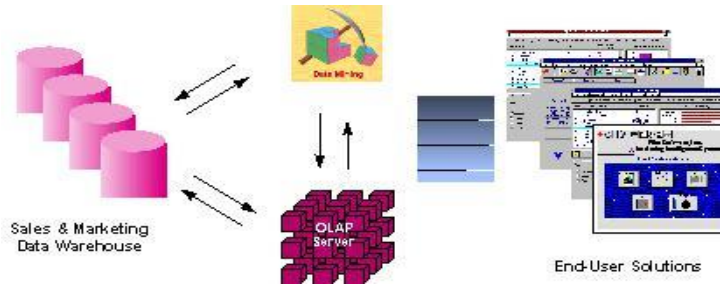


**Figure 1 – Integrated Data Mining Architecture**

The best point to start is a data warehouse containing a combination of internal data tracking all customer contact coupled with external market data about competitor activity. This warehouse can be implemented in a variety of relational database systems.

This design represents a fundamental shift from conventional decision support systems. Rather than simply delivering data to the end user through query and reporting software, the Advanced Analysis Server applies users' business models directly to the warehouse and returns a proactive analysis of the most relevant information. These results enhance the metadata in the OLAP Server by providing a dynamic metadata layer that represents a distilled view of the data. Reporting, visualization, and other analysis tools can then be applied to plan future actions and confirm the impact of those plans.

## Scope of Data Mining

Data mining is termed as searching for valuable business information in a large database and this process requires either sifting through an immense amount of material, or intelligently probing it to find exactly where the value resides. Given databases of sufficient size and quality, data mining technology can generate new business opportunities by providing these capabilities:

- *Automated prediction of trends and behaviors:* Data mining automates the process of finding predictive information in large databases. Questions which used to require an extensive analysis are now quickly answered directly from data. For example: targeted marketing. Data mining uses data on past promotional mailings to identify the targets most likely to maximize return on investment in future mailings
- *Automated discovery of previously unknown patterns:* Data mining tools go through large databases and identify previously hidden patterns in one step. For example the analysis of retail sales data to identify seemingly unrelated products that are often purchased together.

Data mining techniques can yield the benefits of automation on existing software and hardware platforms, and can be implemented on new systems as existing platforms are upgraded and new products developed. When data mining tools are implemented on high performance parallel processing systems, they can analyze massive databases in minutes. Faster processing means that users can automatically experiment with more models to understand complex data. High speed makes it practical for users to analyze huge quantities of data. Larger databases, in turn, yield improved predictions.

## Conclusion

Since data mining is a young discipline with wide and diverse applications, there is still a nontrivial gap between general principles of data mining and domain specific, effective data mining tools for particular applications. A few application domains of Data Mining and Trends in Data Mining which include further efforts towards the exploration of new application areas and new methods for handling complex data types, algorithms scalability, constraint based mining and visualization methods, the integration of data mining with data warehousing and database systems, the standardization of data mining languages, and data privacy protection and security.

## References


[1] Rakesh Agrawal et al.: Fast Discovery of Association Rules. Advances in Knowledge Discovery and Data Mining 1996: 307-328.

[2] Surajit Chaudhuri: Data Mining and Database Systems: Where is the Intersection? Data Engineering Bulletin 21(1) 1998.

[3] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth, From Data mining to knowledge Discovery in Databases, AI Magazine 17(3): Fall 1996, 37-54.

[4] Data Mining Concepts and Techniques – Jiawei Han & Micheline Kamber.

[5] Agrawal, R. and Srikant, R. Fast algorithms for mining association rules. Proc. of Conf. Very Large Data Bases (VLDB'94), pp. 487-499, Santiago, Chile Sept., 1994.

[6] Brin Sergey, Motwani Rajeev, Ullman Jeffrey, D. and Tsur Shalom. Dynamic item set counting and implication rules for market basket data. Proc. of ACM SIGMOD international conference on management of data, vol. 26, issue 2, pp 255-264, 1997.

[7] Bastide Yves. Taouil Rafik. Pasquier Nicolas. Stumme Gerd and Lakhal Lotfi. Mining Frequent Patterns with Counting Inference. In Proc. of ACM SIGKDD, pp 68-75, 2000.

[8] Sun Ken and Bai Fengshan. Mining Weighted Association Rules without Pre assigned Weights. In Proc. Of IEEE Transactions on Knowledge and Data Engineering, Vol. 20, No. 4, pp 489-495, 2008.

[9] Jiawei Han and Michline Kamber . Data Mining Concepts and Techniques. Second edition, The Morgan Kaufmann series in Data Management Syatems, 2006.

[10] Sunita Sarawagi, Shiby Thomas, Rakesh Agrawal: Integrating Association Rule Mining with Relational Database Systems: Alternatives and Implications. Data Mining and Knowledge Discovery 4 (2/3).

[11] John Clear et al.: NonStop SQL/MX Primitives for Knowledge Discovery. KDD 1999: 425-429.