IJREAT International Journal of Research in Engineering & Advanced Technology, Volume 1, Issue 5, Oct-Nov, 2013
ISSN: 2320 - 8791
www.ijreat.org

Data Mining Designing Data Warehousing

SAMYUKTHA.N[1], VENKATESHWARI.R[2] and VICTORIYA.P[3]

SNS College of Engineering, Coimbatore

SNS College of Engineering, Coimbatore

Abstract

Data mining is a combination of database and artificial intelligence technologies. Although the AI field has taken a major dive in the last decade; this new emerging field has shown that AI can add major contributions to existing fields in computer science. In fact, many experts believe that data mining is the third hottest field in the industry behind the Internet, and data warehousing. Data mining is really just the next step in the process of analyzing data. Instead of getting queries on standard or user-specified relationships, data mining goes a step farther by finding meaningful relationships in data. Relationships that were thought to have not existed or ones that give a more insightful view of the data. For example, a computer-generated graph may not give the user any insight, however data mining can find trends in the same data that shows the user more precisely what is going on. Using trends that the end-user would have never thought to query the computer about. Without adding any more data, data mining gives a huge increase in the value added by the database. It allows both technical and non-technical users get better answers, allowing them to make a much more informed decision, saving their companies millions of dollars Data warehousing and on-line analytical processing (OLAP) are essential elements of decision support, which has increasingly become a focus of the database industry. Many commercial products and services are now available, and all of the principal database management system vendors now have offerings in these areas. Decision support places some rather different requirements on database technology compared to traditional on-line transaction processing applications. This paper provides an overview of data warehousing and OLAP technologies, with an emphasis on their new requirements. We describe back end tools for extracting, cleaning and loading data into a data warehouse; multidimensional data models typical of OLAP; front end client tools for querying and data analysis; server extensions for efficient query processing; and tools for metadata management and for managing the warehouse. In addition to surveying the state of the art, this paper also identifies some promising research issues, some of which are related to problems that the database research community has worked on for years, but others are only just beginning to be addressed. This overview is based on a tutorial that the authors presented at the VLDB Conference, 1996.

1. INTRODUCTION

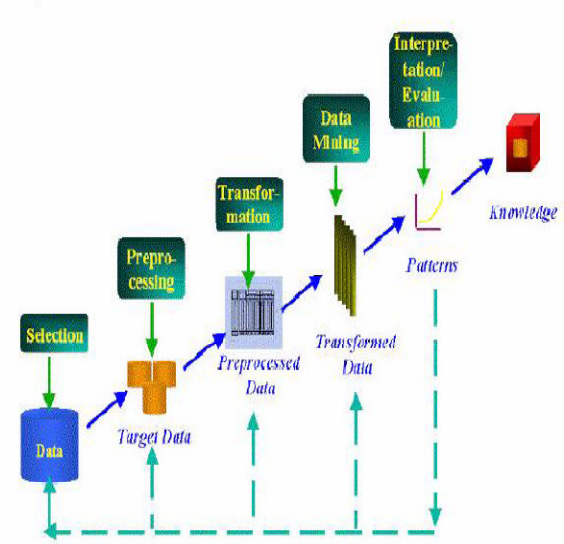
"Data mining is the process of discovering meaningful new correlations, patterns, and trends by sifting through large amounts of data stored in repositories, using pattern recognition technologies as well as statistical and mathematical techniques" (SPSS). However, really data mining turns databases into knowledge bases which is one of the fundamental components of expert systems. Instead of the computer just blindly pulling data from a database, the computer is able to take all the data and interpret it, which is a huge step to make. If it was not for existing AI technologies this field could not have emerged as quickly; if at all. Data warehouse is a computer system designed to give business decision-makers instant access to information. The warehouse copies its data from existing systems like order entry, general ledger, and human resources and stores it for use by executives rather than programmers. Data warehouse users use special software that enables them to create and access information when they need it, as opposed to a reporting schedule defined by the information systems(IS) department.

Databases today can range in size into tera bytes. Within these masses of data lies hidden information of strategic importance. Data mining is the process of uncovering that information. Innovative organizations worldwide are already using data mining to locate and appeal to higher-value customers, to reconfigure their product offerings to increase sales, and to minimize losses due to error or fraud. Data mining is a relatively unique process which extracts information from a database that the user did not know existed. There's the more difficult way to use the results of data mining: getting the user to actually understand what is going on so that they can take action directly.

1) Visualization of the data mining output in a meaningful way, and

2) Allowing the user to interact with the visualization so that simple questions can be answered.

2. STAGES OF DATA MINING PROCESS



3. DATA WAREHOUSING

Data mining potential can be enhanced if the appropriate data has been collected and stored in a data warehouse. Data warehousing is technique making it possible to extract archived operational data and overcome inconsistencies between different legacy data formats.

3.1 Characteristics of Data Warehouse

- *Subject-oriented*: data is organized according to subject instead of application
- *Integrated*: When data resides in many separate applications in the operational environment, encoding of data is often inconsistent.
- *Time- variant*: The data warehouse contains data for comparisons.
- *Non-volatile:* Data are not updated or changed in any way once they enter the data warehouse, but are only loaded and accessed.

3.2 Architecture of Data Mining And Data Warehouse System

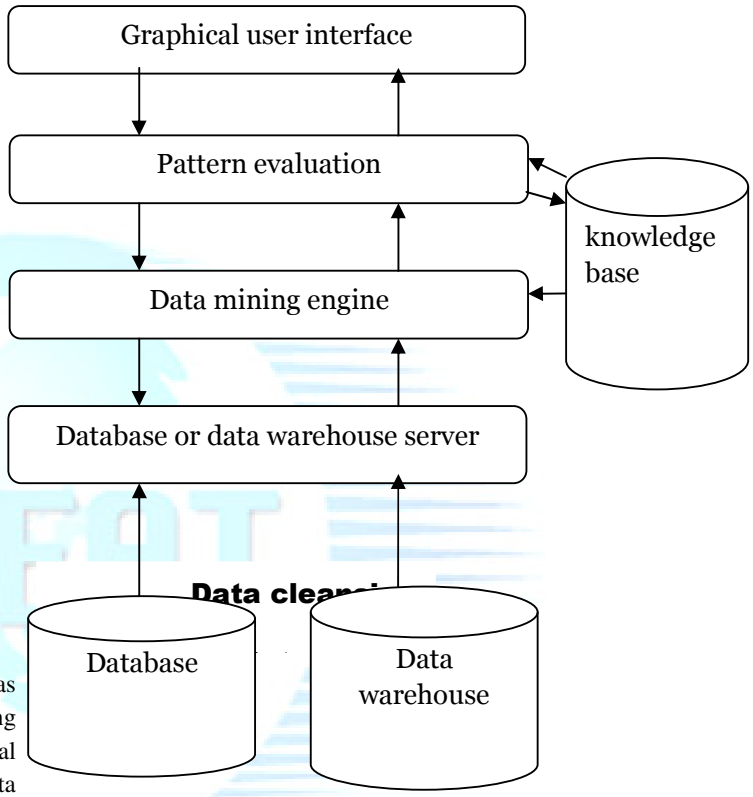
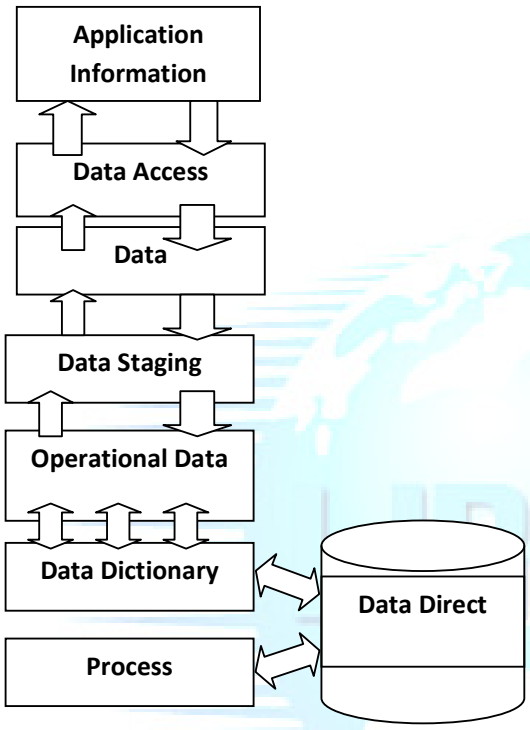


Fig. Data Warehousing



3.3 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.

3.4 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.

Data mining consists of five major elements:

- Extract, transform, and load transaction data onto the data warehouse system.
- Store and manage the data in a multidimensional database system.
- Provide data access to business analysts and information technology professionals.
- od of a backpack being purchased based on a consumer's purchase of sleeping bags and hiking shoes.

4. DATA MINING TECHNIQUES

Classical Techniques

Statistics

By strict definition "statistics" or statistical techniques are not data mining. They were being used long before the term data mining was coined to apply to business applications

Nearest Neighbor

Clustering and the Nearest Neighbor prediction technique are among the oldest techniques used in data mining. Nearest neighbor is a prediction technique that is quite similar to clustering.

Clustering

Clustering is the method by which like records are grouped together. Usually this is done to give the end user a high level view of what is going on in the database. Clustering is sometimes used to mean segmentation.

5. NEXT GENERATION TECHNIQUES

The data mining techniques in this section represent the most often used techniques that have been developed over the last two decades of research. These techniques can be used for either discovering new information within large databases or for building predictive models.

5.1 Decision Trees

A decision tree is a predictive model that, as its name implies, can be viewed as a tree. Specifically each branch of the tree is a classification question and the leaves of the tree are partitions of the dataset with their classification. For instance if we were going to classify customers who churn(don't renew their phone contracts) in the Cellular Telephone Industry a decision tree might look something like that found in Figure.

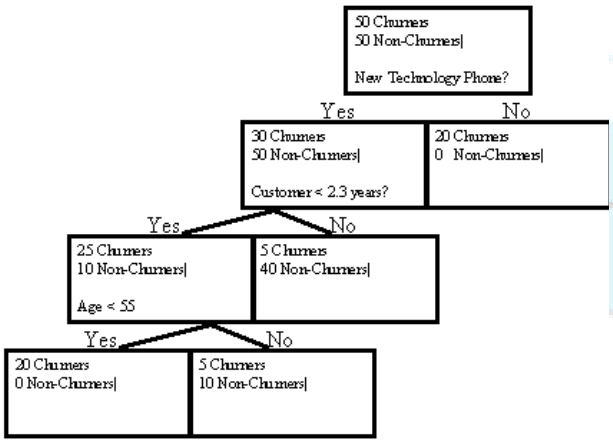


Figure:A decision tree is a predictive model that makes a prediction on the basis of a series of decision much like the game of 20 questions.

We may notice some interesting things about the tree:

- It divides up the data on each branch point without losing any of the data (the number of total records in a given parent node is equal to the sum of the records contained in its two children).
- The number of churners and non-churners is conserved as you move up or down the tree

5.2 Which Technique and When?

Some of the criteria that are important in determining the technique to be used are determined by trial and error. There are definite differences in the types of problems that are most conducive to each technique but the reality of real world data and the dynamic way in which markets, customers and hence the data that represents them is formed means that the data is constantly changing. These dynamics mean that it no longer makes sense to build the "perfect" model on the historical data since whatever was known in the past cannot adequately predict the future because the future is so unlike what has gone before.

6. POTENTIAL APPLICATIONS

Data mining has many and varied fields of application some of which are listed below.

Retail/Marketing

Banking

Insurance and Health Care

Transportation

Medicine

7. CONCLUSION

Data Mining is not a new phenomenon. All large organizations already have data warehouses, but they are just not managing them. The Data Warehousing solution should enhance intelligence in decision-making process of an enterprise. Over the next few years, the growth of data mining is going to be enormous with new products and technologies coming out frequently. In order to get the most out of this period, it is going to be important that data warehousing and mining planners and developers have a clear idea of what they are looking for and then choose strategies and methods that will provide them with performance today and flexibility for tomorrow.