RESEARCH ARTICLE OPEN ACCESS

An Overview of Data Mining and Warehousing – Architecture, Techniques and Applications

P. Nithya, G. Lakshmipriya
Assistant Professor
Department of Computer Science
Srimad Andavan Arts & Science College
Tiruchirappalli
Tamilnadu - India

ABSTRACT

Data mining is the process of extract in patterns from data. As more data is gathered, with the amount of data doubling every three years, data mining is becoming an increasingly important tool to transform these data into information. It is commonly used in a wide range of profiling practices, such as marketing, surveillance, fraud ddetection and scientific discovery. The term data mining has also been used in a related but negative sense, to mean the deliberate searching for apparent but not necessarily representative patterns in large numbers of data. To avoid confusion with the other sense, the terms data dredging and data snooping are often used. The data warehouse contains data from most or all of an organization's operational systems and this data is made consistent. Data in the data warehouse is never over-written or deleted - once committed, the data is static, read-only, and retained for future reporting. Top-down data warehousing, has defined a data warehouse as a centralized repository for the entire Enterprise. Bottom – up data warehousing, is a proponent of an approach to data warehouse design frequently.

***Keywords*:-** Data Dredging, Data Snooping, Integrated, Non-Volatile**.**

I. 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 - 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.

Data mining derives its name from the similarities between searching for valuable business information in a large database. "A warehouse is a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision making process". Data warehousing is essentially what you need to do in order to create a data warehouse, and what you do with it.

II. DATA, INFORMATION, AND KNOWLEDGE

DATA

Data are any facts, numbers, or text that can be processed by a computer. Today, organizations are accumulating vast and growing amounts of data in different formats and different databases. This includes:

- Operational or transactional data such as, sales, cost, inventory, payroll, and accounting.
- Nonoperational data, such as industry sales, forecast data, and macro-economic data.
- Meta data - data about the data itself, such as logical database design or data dictionary definitions

INFORMATION

The patterns, associations, or relationships among all this *data* can provide *information*. For example, analysis of retail point of sale transaction data can yield information on which products are selling and when.

KNOWLEDGE

Information can be converted into ***knowledge*** about historical patterns and future trends. For example, summary Information on retail supermarket sales can be analyzed in light of promotional efforts to provide knowledge of

consumer buying behavior. Thus, a manufacturer or retailer could determine which items are most susceptible to promotional efforts.

III. DATA MINING ARCHITECTURE

Data mining is the process of extracting patterns from data. As more data is gathered, with the amount of data doubling every three years, data mining is becoming an increasingly important tool to transform these data into information. It is commonly used in a wide range of profiling practices, such as marketing, surveillance, fraud detection and scientific discovery.

The term data mining has also been used in a related but negative sense, to mean the deliberate searching for apparent but not necessarily representative patterns in large numbers of data. To avoid confusion with the other sense, the terms data dredging and data snooping are often used. Note, however, that dredging and snooping can be (and sometimes are) used as exploratory tools when developing and clarifying hypotheses.

A primary reason for using data mining is to assist in the analysis of collections of observations of behavior. Such data are vulnerable to collinearity because of unknown interrelations. An unavoidable fact of data mining is that the (sub-) set(s) of data being analyzed may not be representative of the whole domain, and therefore may not contain examples of certain critical relationships and behaviors that exist across other parts of the domain. To address this sort of issue, the analysis may be augmented using experiment-based and other approaches, such as Choice Modelling for human-generated data. In these situations, inherent correlations can be either controlled for, or removed altogether, during the construction of the experimental design.

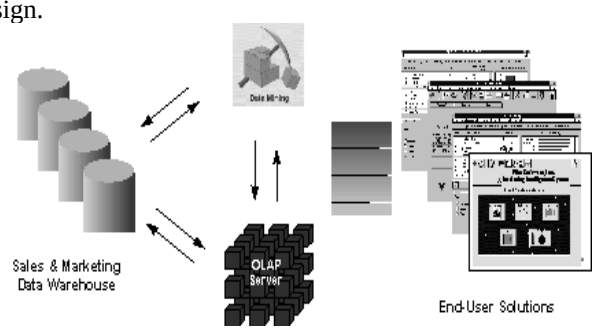

Figure 1 - Integrated Data Mining Architecture

Data mining commonly involves four classes of tasks.

- Classification - Arranges the data into predefined groups. For example an email program might attempt to classify an email as legitimate or spam. Common algorithms include Decision Tree Learning, Nearest Neighbor, Naive Bayesian Classification and Neural Network.
- Clustering - Is like classification but the groups are not predefined, so the algorithm will try to group similar items together.
- Regression - Attempts to find a function which models the data with the least error.
- Association rule learning - Searches for relationships between variables. For example a supermarket might gather data on customer purchasing habits. Using association rule learning, the supermarket can determine which products are frequently bought together and use this information for marketing purposes. This is sometimes referred to as "market basket analysis".

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.
- Analyze the data by application software.
- Present the data in a useful format, such as a graph or table.

IV. DATAMINING TECHNIQUE

"Pattern mining" is a data mining technique that involves finding existing patterns in data. In this context *patterns* often means association rules. The original motivation for searching association rules came from the desire to analyze supermarket transaction data, that is, to examine customer behavior in terms of the purchased products.

"Subject-based data mining" is a data mining technique involving the search for associations between individuals in data. In the context of combatting terrorism, the National Research Council provides the following definition: *"Subject-based data mining* uses an initiating individual or other datum that is considered, based on other information, to be of high interest, and the goal is to determine what other persons or financial transactions or movements, etc., are related to that initiating datum.

The most commonly used techniques in data mining are:

- Artificial neural networks: Non-linear predictive models that learn through training and resemble biological neural networks in structure.
- Decision trees. Tree-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) .

- Genetic algorithms. Optimization techniques that use processes such as genetic combination, mutation, and natural selection in a design based on the concepts of evolution.
- 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. Sometimes called the k-nearest neighbor technique.
- Rule induction. The extraction of useful if-then rules from data based on statistical significance.

V. APPLICATIONS

Businesses employing data mining may see a return on investment, but also they recognize that the number of predictive models can quickly become very large. Rather than one model to predict which customers will churn, a business could build a separate model for each region and customer type. It may also want to determine which customers are going to be profitable over a window of time and only send the offers to those that are likely to be profitable. In order to maintain this quantity of models, they need to manage model versions and move to *automated data mining*.

Data mining can also be helpful to human-resources departments in identifying the characteristics of their most successful employees.

Another example of data mining, often called the market basket analysis, relates to its use in retail sales.

Market basket analysis has also been used to identify the purchase patterns of the Alpha consumer. Alpha Consumers are people that play a key role in connecting with the concept behind a product, then adopting that product, and finally validating it for the rest of society.

In recent years, data mining has been widely used in area of science and engineering, such as bioinformatics, genetics, medicine, education and electrical power engineering.

In the area of study on human genetics, the important goal is to understand the mapping relationship between the inter-individual variation in human DNA sequences and variability in disease susceptibility.

In the area of electrical power engineering, data mining techniques have been widely used for condition monitoring of high voltage electrical equipment. The purpose of condition monitoring is to obtain valuable information on the insulation's health status of the equipment.

Data mining techniques have also been applied for dissolved gas analysis (DGA) on power transformers. DGA, as a diagnostics for power transformer, has been available for many years

A fourth area of application for data mining in science/engineering is within educational research, where data mining has been used to study the factors leading students to choose to engage in behaviors which reduce their learning and to understand the factors influencing university student retention.

Other examples of applying data mining technique applications are biomedical data facilitated by domain on topologies, mining clinical trial data, traffic analysis using SOM, etcetera.

Limitations of Data Mining

While data mining products can be very powerful tools, they are not self-sufficient applications. To be successful, data mining requires skilled technical and analytical specialists who can structure the analysis and interpret the output that is created. Consequently, the limitations of data mining are primarily data or personnel related, rather than technology-related. Although data mining can help reveal patterns and relationships, it does not tell the user the value or significance of these patterns. These types of determinations must be made by the user.

VI. DATA WAREHOUSING

The data warehousing market consists of tools, technologies, and methodologies that allow for the construction, usage, management, and maintenance of the hardware and software used for a data warehouse, as well as

the actual data itself. Surveys indicate Data Warehousing will be the single largest. Data warehousing is currently a $28 Billion market (Source: Data Warehousing Institute) and we estimate 20% growth per annum through at least 2002

"A warehouse is a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision making process".

Subject Oriented

Data that gives information about a particular subject instead of about a company's ongoing operations.

Integrated

Data that is gathered into the data warehouse from a variety of sources and merged into a coherent whole.

Time-variant

All data in the data warehouse is identified with a particular time period.

Non-volatile

Data is stable in a data warehouse. More data is added but data is never removed. This enables management to gain a consistent picture of the business.

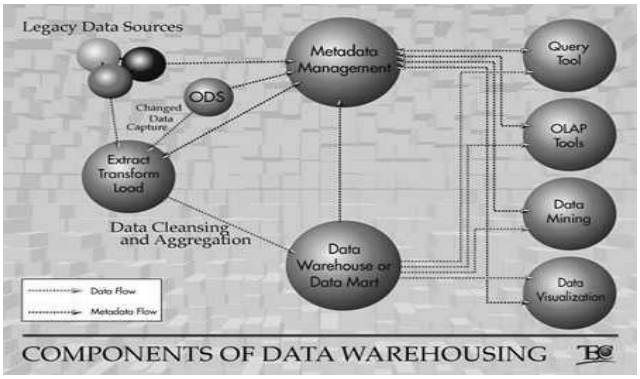

Figure – 2. Data Warehousing Components

Benefits

Some of the benefits that a data warehouse provides are as follows

- A data warehouse provides a common data model for all data of interest regardless of the data's source. This makes it easier to report and analyze information than it would be if multiple data models were used to retrieve information such as sales invoices, order receipts, general ledger charges, etc.
- Prior to loading data into the data warehouse, inconsistencies are identified and resolved. This greatly simplifies reporting and analysis.
- Information in the data warehouse is under the control of data warehouse users so that, even if the source system data is purged over time, the information in the warehouse can be stored safely for extended periods of time.
- Because they are separate from operational systems, data warehouses provide retrieval of data without slowing down operational systems.
- Data warehouses can work in conjunction with and, hence, enhance the value of operational business applications, notably customer relationship management (CRM) systems.
- Data warehouses facilitate decision support system applications such as trend reports (e.g., the items with the most sales in a particular area within the last two years), exception reports, and reports that show actual performance versus goals.

Disadvantages

There are also disadvantages to using a data warehouse. Some of them are:

- Data warehouses are not the optimal environment for unstructured data.
- Because data must be extracted, transformed and loaded into the warehouse, there is an element of latency in data warehouse data.
- Over their life, data warehouses can have high costs. Maintenance costs are high.
- Data warehouses can get outdated relatively quickly. There is a cost of delivering suboptimal information to the organization.
- There is often a fine line between data warehouses and operational systems. Duplicate, expensive functionality may be developed. Or, functionality may be developed in the data warehouse that, in retrospect, should have been developed in the operational systems and vice versa.

SAMPLE APPLICATIONS

Some of the applications data warehousing can be used for are:

- Credit card churn analysis
- Insurance fraud analysis
- Call record analysis
- Logistics management.

VII. CONCLUSION

Comprehensive data warehouses that integrate operational data with customer, supplier, and market information have resulted in an explosion of information. Competition requires timely and sophisticated analysis on an integrated view of the data. However, there is a growing gap between more powerful storage and retrieval systems and the users' ability to effectively analyze and act on the information they contain. Both relational and OLAP technologies have tremendous capabilities for navigating massive data warehouses, but brute force navigation of data is not enough. A new technological leap is needed to structure and prioritize information for specific end-user problems. The data mining tools can make this leap. Quantifiable business benefits have been proven through the integration of data mining with current information systems,

and new products are on the horizon that will bring this integration to an even wider audience of users.

REFERENCES

[1] The Morgan Kaufmann Series in Data Management Systems, Jim Gray, Series Editor Morgan Kaufmann Publishers, March 2006. ISBN 1-55860-901-6

[2] Knowledge warehouse: an architectural integration of knowledge management, decision support, artificial intelligence and data warehousing - HR Nemati, DM Steiger, LS Iyer, RT Herschel - Decision Support Systems, 2002 – Elsevier

[3] An empirical investigation of the factors affecting data warehousing success BH Wixom, HJ Watson - MIS quarterly, 2001 – JSTOR

[4] A review of data mining techniques SJ Lee, K Siau - Industrial Management & Data Systems, 2001 – emeraldinsight.com

[5 Enhancing usability testing through datamining techniques: A novel approach to detecting usability problem patterns for a context of use MP González, J Lorés, A Granollers - Information and software technology, 2008 – Elsevier

[6] A survey of clustering data mining techniques P Berkhin - Grouping multidimensional data, 2006 – Springer

[7] Data mining techniques M Zaki, L Wong - Selected Topics in Post-Genome Knowledge …, 2003 – ims.nus.edu.sg

[8] Towards a framework for evaluating investments in data warehousing A Counihan, P Finnegan… - … Systems Journal, 2002 - Wiley Online Library.

[9] Fundamentals of spatial data warehousing for geographic knowledge discovery Y Bédard, T Merrett, J Han - … data mining and knowledge …, 2001 - books.google.com