Overview of Data Warehouse and Data Mining

Author: Mrs. Rutuja Tendulkar
Lecturer, V.P.M's Polytechnic, Thane

Abstract: Today in organizations, the developments in the transaction processing technology requires that, amount and rate of data capture should match the speed of processing of the data into information which can be utilized for decision making. A data warehouse is a subject-oriented, integrated, time-variant and non-volatile collection of data that is required for decision making process. Data mining involves the use of various data analysis tools to discover new facts, valid patterns and relationships in large data sets. Data mining also includes analysis and prediction for the data. Data mining helps in extracting meaningful new patterns that cannot be found just by querying or processing data or metadata in the data warehouse. This paper includes need for data warehousing and data mining, how data warehousing and mining helps decision making systems, Knowledge Discovery process and various techniques involve in data mining.

Keywords: Data, Information, Decision, Warehousing, Mining

Data warehousing:

Large amount of operational data are routinely collected and stored away in the archives of many organizations. To take a simple example, the railway reservation system has been operational for over a decade and large amount of data is generated each day on train bookings. Much of this data is probably archived for audit purposes. This archived operational data can be effectively used for tactical strategic management of the railways. For example, by analyzing the reservation data it would be possible to find out traffic patterns in various sectors and use it to add or remove bogies in certain trains, to decide on the mix of various classes of accommodation, etc.For this analysis building a data warehouse is an effective solution. Data warehouse is a storage area for processed and integrated data across different sources which will be both operational data and external data. Data warehouses offer organizations the ability to gather and store enterprise information in a single conceptual enterprise repository. It allows its users to extract required data for business analysis and strategic decision making. One can also define a warehouse as a copy of transaction data specifically structured for query and analysis. It is a repository of information, integrated from several operational databases. Data warehouses store large amount of data which can be frequently used by decision support system. It is maintained separately from the organizations operational database. They are relatively static with only infrequent updates. The most effective advantages of data warehousing is high speed of data processing and summarized data.

Characteristics of data warehouse:

- Subject oriented: A data warehouse is organized around major subjects such as customer, products, sales; etc.Data is organized according to subject instead of application. For example, an insurance company using a data warehouse would organize their data by customer, premium and claim instead of by different product like auto, life; etc.The data organized by subject obtains only the information necessary for the decision support processing.
- Integrated: A data warehouse is usually constructed by integrating multiple, heterogeneous sources such as relational databases, flat files, and OLTP file. When data resides in many separate applications in the operational environment, the encoding of data is often inconsistent. When data is moved from operational environment into the data warehouse, they assume a consistent coding convention. Data cleaning and data integration techniques are applied to maintain consistency in naming convention, measures of variables, encoding structure and physical attributes.
- Nonvolatile: A data warehouse is always a physically separate store of data, which is transformed from the application data found in the appropriate environment. Due to this separation, data warehouses do not require transaction processing, recovery, concurrency control, etc.The data is not updated or changed in any way once they enter the data warehouse, but are only loaded, refreshed and accessed for queries.
- Time variant: Data is stored in data warehouse to provide a historical perspective. Every key structure in the data warehouse contains, implicitly or explicitly, an element of time. The data warehouse contains a place for sorting data that are 5 to 10 years old, or older, to be used for comparisons, trends and forecasting.

Database vs. data warehouse:

Database is a collection of related information stored in a structured form in terms of table so that it makes easier insertion, deletion and manipulation of data. Database consists of tables that contain attributes. Whereas a data warehouse is a database system optimized for reporting and analysis. It generally refers to the combination of many different databases across entire enterprise. Once the data entered in the data warehouse, it can be then only loaded, refreshed and accessed for queries.

Data mining:

It is a process of extracting hidden predictive information from large databases. It is a powerful new technology to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. For a commercial business, the discovery of previously unknown statistical patterns or trends can provide valuable insight into the function and environment of their

organization. Data-mining techniques can generally be grouped into two categories: predictive method and descriptive method.

Descriptive method: It a method of finding human interpretable patterns that describe the data. Data mining in this case is useful to group together similar documents returned by search engine according to their context.

Predictive method: In this method, we can use some variables to predict unknown or future values of other variable. It is used to predict whether a newly arrived customer will spend more than 100$ at a department store.

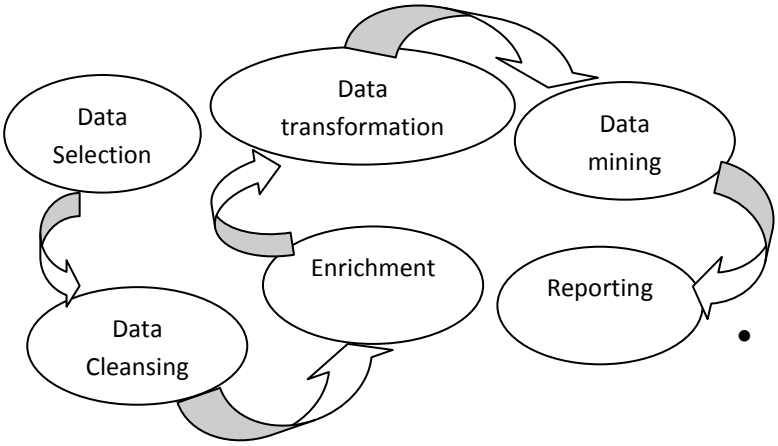
Data-mining techniques:

The following list describes many data-mining techniques in use today. Each of these techniques exists in several variations and can be applied to one or more of the categories above.

- Regression modeling—This technique applies standard statistics to data to prove or disprove a hypothesis. One example of this is linear regression, in which variables are measured against a standard or target variable path over time. A second example is logistic regression, where the probability of an event is predicted based on known values in correlation with the occurrence of prior similar events.
- Visualization—This technique builds multidimensional graphs to allow a data analyst to decipher trends, patterns, or relationships.
- Correlation—This technique identifies relationships between two or more variables in a data group.
- Variance analysis—This is a statistical technique to identify differences in mean values between a target or known variable and nondependent variables or variable groups.
- Discriminate analysis—This is a classification technique used to identify or "discriminate" the factors leading to membership within a grouping.
- Forecasting—Forecasting techniques predict variable outcomes based on the known outcomes of past events.
- Cluster analysis—This technique reduces data instances to cluster groupings and then analyzes the attributes displayed by each group.
- Decision trees—Decision trees separate data based on sets of rules that can be described in "if-then-else" language.
- Neural networks—Neural networks are data models that are meant to simulate cognitive functions. These techniques "learn" with each iteration through the data, allowing for greater flexibility in the discovery of patterns and trends.

Data mining as a part of Knowledge discovery in database:

Data mining addresses inductive knowledge which discovers new rules and patterns from the supplied data.It comprises six phases such as data selection, data cleansing, enrichment, data transformation or encoding, data mining and the reporting and display of the discovered information.



KDD process: Consider a transaction database maintained by a specialty consumer goods retailer. Client data includes customer name, zip code, phone number, data of purchase, item code, price, quantity and total amount.KDD process can be applied to this database to discover variety of knowledge.

- Data selection: Selecting a data set, or focusing on a subset of variables or data samples, on which discovery is to be performed. Data about specific item or categories of items, or from stores in a specific region or area of the country, may be selected.
- Data cleansing: It checks and resolves data conflicts, outliers, noisy, erroneous, missing data and ambiguity. In this they may correct invalid zip codes or eliminate records with incorrect phone prefixes.
- Enrichment: Enhances the data with additional sources of information. For example, with client name and phone numbers, new information about the client such as age, income and credit rating can be appended to each record.
- Data transformation and encoding: Data is transformed or consolidated into forms appropriate for mining, by performing summary, or aggregation, operations.It is done to reduce the amount of data. For example, item codes may be grouped in terms of product categories into audio, video, supplies, accessories and so on.
- Data mining: It is a process where intelligent methods are applied to extracts meaningful new patterns. It searches for patterns of interest in a particular representational form or a set of such representations, including classification rules or trees, regression, clustering, sequence modeling and so on.
- Reporting: The results of data mining may be reported in a variety of formats, such as listings, graphic outputs, summary tables or visualizations.

Conclusion:

Organizations today are under tremendous pressure to compete in an environment of tight deadlines and reduced profits. Business processes that require data to be extracted and manipulated prior to use will no longer be acceptable. Instead, enterprises need rapid decision support based on the analysis and forecasting of predictive behavior. Data-warehousing and data-mining techniques provide this capability.

References:

http://www.executionmih.com/data-mining/kdd-process-preparation-evaluation.php

http://articles.techrepublic.com.com/5100-10878_11-1045046.html

www.youtube.com/video

Book-Fundamentals of Database Systems-Fourth edition by Remez Elmasri & Shamkant Navathe

Book-Data mining techniques by Arun K. Pujari