DATA WAREHOUSING AND DATA MINING

Abilash Sasi
Vth Semester, Raipur Institute of Technology, Raipur.
Email: abilashsasi1@yahoo.co.in

Abhinandan Majumdar
VIth Semester, National Institute of Technology Karnataka, Surathkal.
Email: abhi_bsp_nitk@yahoo.co.in

ABSTRACT

The increasing processing power and sophistication of analytical tools and techniques have resulted in the development of what are known as Data Warehouses. These Data Warehouses provide storage, functionality, responsiveness to queries beyond the capabilities of transaction oriented databases. At present there is a great need to provide decision makers from middle management upward with information at the correct level of detail to support decision making. Both Data Warehousing and Data Mining provide this facility.

Data Mining or Knowledge Discovery in databases has been popularly recognized as an important research issue with broad applications. It refers to the mining or discovery of new information in terms of patterns or rules from vast amounts of data. We provide a comprehensive survey in database perspective, on the database techniques developed recently also focusing on Data Mining methods i.e. class description, association, classification, prediction, clustering and time series analysis. Data Mining Research is focused on high quality and scalable clustering methods for large databases and multidimensional databases.

Clustering Analysis is to identify clusters embedded in the data, where a cluster is a collection of data objects that are similar to one another. Clustering methods of machine learning place great importance on the utility of conceptual descriptions, which logically or probabilistically express patterns found in clusters. A good clustering technique produces high quality clusters to ensure that the inter-cluster similarity is low and intra-cluster similarity is high. Conceptual descriptions are important for cluster interpretation, inference tasks such as pattern completion and problem solving, and for data compression, memory management and runtime-efficiency enhancements.

DATA WAREHOUSES

Introduction

Data warehousing is a collection of *decision support* technologies, aimed at enabling the *knowledge worker* (executive, manager, and analyst) to make better and faster decisions. Data warehousing technologies have been successfully deployed in many industries: manufacturing (for order shipment and customer support), retail (for user profiling and inventory management), financial services (for claims analysis, risk analysis, credit card analysis, and fraud detection), transportation (for fleet management), telecommunications (for call analysis and fraud detection), and utilities (for power usage analysis), and healthcare (for outcomes analysis). This paper presents a roadmap of data warehousing technologies, focusing on the special requirements that data warehouses place on database management systems (DBMSs).

Definition

A data warehouse is a "subject-oriented, integrated, time-varying, non-volatile collection of data that is used primarily in organizational decision making." as defined by W.H.Inmon. Data warehouses provide access to data for complex analysis, knowledge discovery and decision making. It is supposed to be a repository for the useful data collected by those business systems. Data warehousing is the process of making our operational data available to our business managers and decision support applications. Several types of applications – OLAP, DSS and Data mining applications are supported which are discussed below.

OLAP (online analytical processing) is a term used to describe the analysis of complex data from the data warehouses. These tools use distributed computing capabilities for analyses that require more storage and processing power that can be economically and efficiently located on an individual desktop.

DSS (decision-support systems) also known as EIS (executive information systems) support an organization's leading decision makers with higher level data for complex and important decisions. Data Mining (which we will discuss later) is used for knowledge discovery, the process for searching data for unanticipated new knowledge.

Traditional databases support online transaction processing (OLTP), which includes insertions, deletions and updates, while also supporting query requirements. These are optimized to process queries that may touch a small part of the database and transactions that deal with insertions or updates of a few tuples per relation to process. Thus they cannot be optimized for OLAP, DSS, or data mining. By contrast, data warehouses are designed precisely to support efficient extraction processing and presentation for analytic and decision-making process. In comparison to traditional databases, data warehouses generally contain very large amounts of data from multiple sources that may include databases from different data models and sometimes files acquired from independent systems and platforms.

Architecture of Data Warehouse

We can describe data warehousing as "a collection of decision support technologies, aimed at enabling the knowledge worker (executive, manager, and analyst) to make better and faster decisions." Figure 1 gives an overview of the conceptual structure of data warehouses. It shows entire data warehousing process. This includes possible cleaning and reformatting of data before its warehousing. At the back end of the process, OLAP, data-mining and DSS may generate new relevant information such as rules; this information is shown in figure going back to the warehouse. The figure also shows that data sources may include files.

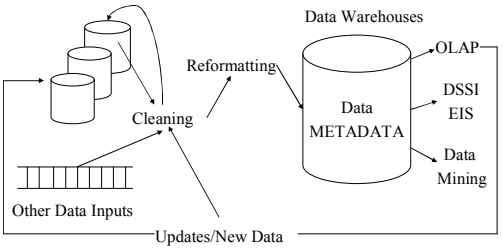

Figure1.Conceptual Structure of data-warehousing process.

Data warehouses have the following distinctive characteristics.

- multidimensional conceptual view
- generic dimensionality
- unlimited dimensions and aggregation levels
- unrestricted cross-dimensional operations
- dynamic sparse matrix handling
- client-server architecture
- multi-user support
- accessibility
- transparency
- intuitive data manipulation
- consistent report performance
- flexible reporting

Data Modeling for Data Warehouse

A popular conceptual model that influences the front-end tools, database design, and the query engines for OLAP is the *multidimensional* view of data in the warehouse. In a multidimensional data model, there is a set of *numeric measures* that are the objects of analysis. Examples of such measures are sales, budget, revenue, inventory, ROI (return on investment). Each of the numeric measures depends on a set of *dimensions,* which provide the context for the measure. For example, the dimensions associated with a sale amount can be the city, product name, and the date when the sale was made. The dimensions together are assumed to *uniquely* determine the measure. Thus, the multidimensional data views a measure as a value in the multidimensional space of dimensions. Each dimension is described by a set of attributes. For example, the Product dimension may consist of four attributes: the category and the industry of the product, year of its introduction, and the average profit margin. For example, the soda Surge belongs to the category beverage and the food industry, was introduced in 1996, and may have an average profit margin of 80%. The attributes of a dimension may be related via a hierarchy of relationships. In the above example, the product name is related to its category and the industry attribute through such a hierarchical relationship.

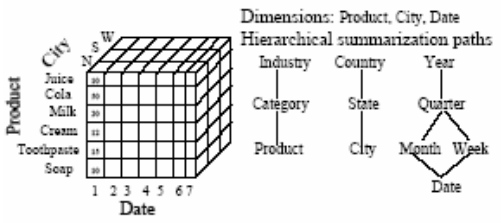

Figure 2. Multidimensional data

Another distinctive feature of the conceptual model for OLAP is its stress on *aggregation* of measures by one or more dimensions as one of the key operations; e.g., computing and ranking the *total* sales by each county (or by each year). Other popular operations include *comparing* two measures (e.g., sales and budget) aggregated by the same dimensions. Time is a dimension that is

of particular significance to decision support (e.g., trend analysis). Often, it is desirable to have built-in knowledge of calendars and other aspects of the time dimension.

The multidimensional storage model involves two types of tables: dimension tables and fact tables. A *dimensional table* consists of tuples of attributes of the dimension. A *fact table* can thought of as having tuples, one per a recorded fact. This fact contains some measured or observed variable(s) and identifies it (them) with pointers to dimension tables. The fact table contains the data, and the dimensions identify each tuples in that data.

Two common multidimensional schemas are the *star schema* and *the snowflake schema.* The star schema consists of a single fact table and a single table for each dimension. Each tuple in the fact table consists of a pointer (foreign key – often uses a generated key for efficiency) to each of the dimensions that provide its multidimensional coordinates, and stores the numeric measures for those coordinates. Each dimension table consists of columns that correspond to attributes of the dimension. Figure 3 shows an example of a star schema

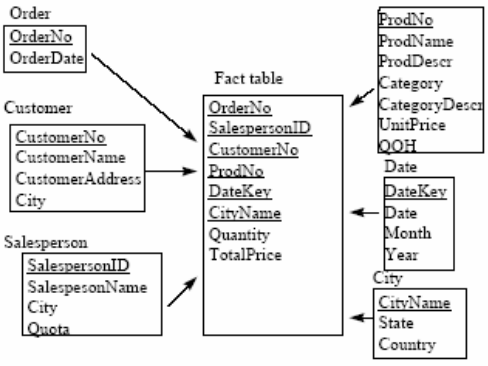

Figure 3. A Star Schema.

Star schemas do not explicitly provide support for attribute hierarchies. *Snowflake schemas* provide a refinement of star schemas where the dimensional hierarchy is explicitly represented by normalizing the dimension tables, as shown in Figure 4. This leads to advantages in maintaining the dimension tables. However, the denormalized structure of the dimensional tables in star schemas may be more appropriate for browsing the dimensions. *Fact constellations* are examples of more complex structures in which multiple fact tables share dimensional tables. For example, projected expense and the actual expense may form a fact constellation since they share many dimensions.

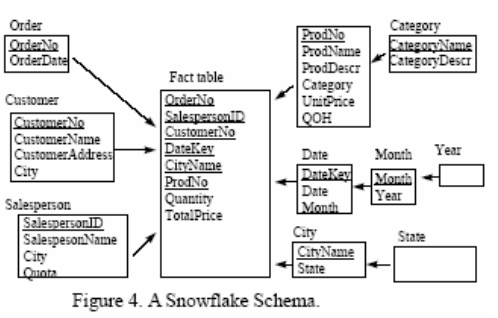

Figure 4. A Snowflake Schema.

Problems in Data Warehouses

Some significant operational issues arise with data warehousing: construction, administration, and quality control. Project Management – the design, construction, and implementation of the warehouse – is an important and challenging consideration that should not be underestimated. The building of an enterprise-wide data warehouse in a large organization is a major undertaking, potentially taking years from conceptualization to implementation.

A significant issue in data-warehousing is the quality control of data. Both quality and consistency of data are major concerns. Although a data passes through a cleaning fraction during acquisition, quality and consistency remain significant concern for the data administrator.

The management of data warehouses also presents new challenges. Detecting runaway queries and managing and scheduling resources are problems that are important but have not been well solved. Some work has been done on the logical correctness of incrementally updating materialized views, but the performance, scalability, and recoverability properties of these techniques have not been investigated. In particular, failure and checkpointing issues in load and refresh in the presence of many indices and materialized views need further research. The adaptation and use of workflow technology might help, but this needs further investigation.

Open Issues in Data Warehouses

Data warehousing as an active research area is likely to see increased research activity in the near future as warehouses and data marts proliferate. Old problems will receive new emphasis; for example, data cleaning, indexing, partitioning, and views could receive renewed attention.

Academic research into data warehousing technologies will likely focus on automating aspects of the warehouse that currently require

significant manual intervention, such as the data acquisitions, data quality management, selection and construction of appropriate access paths and structures, self-maintainability, functionality and performance optimization. Incorporation of domain and business rules appropriately into the warehouse creation and maintenance process may take it intelligent, relevant, and self-governing.

DATA MINING

Introduction

Data mining, *the extraction of hidden predictive information from large databases*, is a powerful new technology with great potential 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. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were too time-consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.

Definition

Data mining is the process of discovering interesting knowledge such as patterns, associations, changes, anomalies and significant structures from large amount of data stored in databases, data warehouses or other information repositories. In a nutshell, it refers to the mining or discovery of new information in terms of patterns or rules from vast amounts of data.

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

Data mining has been popularly treated as a synonym of *knowledge discovery in databases*, although some researchers view data mining as an essential step of knowledge discovery. In general a knowledge discovery process consists of an iterative sequence of following steps.

- **Data Cleaning**: This handles noisy, erroneous, missing or irrelevant data.
- **Data Integration**: where multiple, heterogeneous data sources may be integrated into one.
- **Data Selection:** where data relevant to the analytic task are retrieved from the database.
- **Data Transformation:** where data are transformed and consolidated into forms appropriate for mining by performing summary or aggregation operations.
- **Data Mining:** which is an essential process where intelligent methods are used to extract data patterns.
- **Pattern Evaluation:** This is to identify truly interesting patterns representing knowledge based on some interesting measures.
- **Knowledge Presentation:** where visualization and knowledge representation techniques are used to present the mined knowledge to the user.

Data Mining v/s Data Warehousing

The goal of data warehousing is to support decision making with data. Data mining can be used in conjunction with a data warehouse to help with certain types of decisions. Data mining can be applied to operational databases with individual transactions. To make data mining more efficient, the data warehouse should have an aggregated or summarized collection of data. Data mining helps in extracting meaningful new patterns that cannot be found necessarily by merely querying or processing data or metadata in the data warehouse. Also data mining tools should be designed to facilitate their use in conjunction with data warehouses. In fact, for very large databases running into terabytes of data, successful use of data mining applications will depend first on the construction of a data warehouses.

Goals of Data Mining

Data mining is typically carried out with some end goals or applications. These goals come under following classes: prediction, identification, classification and optimization.

- **Prediction:** Data mining can show how certain attributes within the data will behave in future. Examples predictive data mining include the analysis of buying transactions to predict what consumers will buy under certain discounts, how much sales volume a store would generate in a given period, and whether deleting a product line would yield more profits. In such applications, business logic is coupled with data mining.

- **Identification:** Data patterns can be used to identify the existence of an item, an event, or an activity. For example, intruders trying to break a system may be identified by the programs executed, file accessed, and CPU time per session. The area known as authentication is a form of identification. It ascertains whether a user is indeed a specific user or one from an authorized class, and involves a comparison of parameters or images or signals against a database.

- **Classification:** Data mining can partition the data so that different classes or categories can be identified based on combinations of parameters. For example, customers in a supermarket can be categorized into discount-seeking shopkeepers, shoppers in a rush, loyal regular shoppers, shoppers attached to name brands and infrequent shoppers. This classification may be used in different analyses of customer buying transactions as a post mining activity. Such categorization may be used to encode the data appropriately before subjecting it to further data mining.

- **Optimization:** One eventual goal of data mining may be to optimize the use of limited resources such as time, space, money, or materials and to maximize output variables s such as sales or profits under a given set of constraints. As such, this goal of data mining resembles the objective function used in operations research problems that deals with optimization under constraints.

Classifications of Data Mining

The term "knowledge" is very broadly interpreted as involving some degree of intelligence. Knowledge is often classified as inductive versus deductive. *Deductive knowledge* deduces new information based on applying pre-specified logical rules of deduction on the given data. Data mining addresses *inductive knowledge* which discovers new rules and patterns from supplied data. Knowledge discovered during data mining can be classified as:

- **Class Description:** *Class description* provides a succinct summarization of a collection of data and distinguishes from others. The summarization of a collection of data is called class *characterization;* whereas the comparison of two or more collections of data is called class comparison or discrimination. Class discrimination should not only cover its summary properties such as count, sum, average but also its properties on data dispersion such as variance, quartiles, etc. For example, class description can be used to compare European versus Asian sales of a company, identify the important factors which discriminate two classes, and present a summarized overview.

- **Association:** Association is the discovery of *association relationships* or *correlations* among a set of items. They are often expressed in the rule from showing attribute-value conditions that occur frequently together in a given set of data. An association rule in the form of X=>Y is interpreted as "databases tuples that satisfy X is likely to satisfy Y". Association analysis is widely used in transaction data marketing, catalog design, and other business decision making process. Substantial research has been performed recently on association analysis with efficient algorithms proposed, including the level-wise Apriori search, mining multiple-level, multidimensional associations, mining associations for numerical, categorical, and interval data, meta-pattern directed or constraint based mining and mining correlations.

- **Classification:** Classification analyses a set of training data (i.e., a set of objects whose class label is chosen) and constructs a model for each class based on the features in the data. A *decision tree* or a set of *classification rules* is generated by such a classification process, which can be used for better understanding of each class in database and for classification of future data. For example one may classify diseases and help predict the kind of diseases based on symptoms of patients.

- **Prediction:** This mining function predicts the possible values of some data or value distribution of certain objects in a set of objects. It involves finding of the set of attributes relevant to the attribute of interest (e.g., by some statistical analysis) and predicting the value distribution of similar employees based on the set of data similar to the selected object(s). For example, an employee's potential salary can be predicted based on the salary distribution of the similar employees in the company. Usually, regression analysis, generalized linear model, correlation analysis and decision trees are useful tools in quality predictions. Genetic algorithms and neural network models are also popularly used in prediction.

- **Clustering:** Clustering analysis is to identify clusters embedded in the data, where the cluster is a collection of data objects that are "similar" to one another. Similarity can be specified by the distance functions, specified by users or experts. A good clustering technique produces high quality clusters to ensure that the inter-cluster similarity is low and the intra-cluster similarity is high. For example, one may cluster the houses in an area according to their house category, floor area, and geographic locations. Data mining research has been focused on high quality and scalable clustering methods for large databases and multi dimensional databases.

- **Time-series analysis:** Time-series analysis is to analyze large set of time-series data to find certain regularities and interesting characteristics, including search for similar sequences or subsequences, mining sequential pattern, periodicities, trends and deviations. For example, one may predict the trend of the stock values for a company based on its stock history, business situation, competitors' performance, and current market.

CLUSTERING DATA MINING TECHNIQUE

Introduction

Clustering is a division of data into groups of similar objects. Representing the data by fewer clusters necessarily loses certain fine details, but achieves simplification. It models data by its clusters. Data modeling puts clustering in a historical perspective rooted in mathematics, statistics, and numerical analysis. From a machine learning perspective clusters correspond to hidden patterns, the search for clusters is unsupervised learning, and the resulting system represents a data concept. From a practical perspective clustering plays an outstanding role in data mining applications such as scientific data exploration, information retrieval and text mining, spatial database applications, Web analysis, CRM, marketing, medical diagnostics, computational biology, and many others.

Definition

Clustering is a division of data into groups of similar objects. Each group, called cluster, consists of objects that are similar between themselves and dissimilar to objects of other groups. Representing data by fewer clusters necessarily loses certain fine details (akin to lossy data compression), but achieves simplification. It represents many data objects by few clusters, and hence, it models data by its clusters. Data modeling puts clustering in a historical perspective rooted in mathematics, statistics, and numerical analysis. From a machine learning perspective, clusters correspond to hidden patterns, the search for clusters is unsupervised learning, and the resulting system represents a data concept. Therefore, clustering is unsupervised learning of a hidden data concept. Data mining deals with large databases that impose on clustering analysis additional severe computational requirements. These challenges led to the emergence of powerful broadly applicable data mining clustering methods explained below.

Classification of Clustering Algorithms

Basically clustering techniques are broadly divided in *hierarchical* and *partitioning.*

- **Hierarchical Method**
  - Agglomerative
  - Divisive
- **Partition Clustering**
  - Relocation Algorithms
  - Probabilistic Clustering
  - K mediods methods
  - K means methods

Hierarchical clustering is further subdivided into *agglomerative* and *divisive*. While hierarchical algorithms build clusters gradually (as crystals are grown), partitioning algorithms learn clusters directly. In doing so, they either try to discover clusters by iteratively relocating points between subsets, or try to identify clusters as areas highly populated with data. We will explain both these methods below.

Hierarchical Method

Hierarchical clustering builds a cluster hierarchy or, in other words, a tree of clusters, also known as a *dendrogram.* Every cluster node contains child clusters; sibling clusters partition the points covered by their common parent. Such an approach allows exploring data on different levels of granularity. Hierarchical clustering methods are categorized into

- agglomerative (bottom-up)
- divisive (top-down)

An **agglomerative clustering** starts with one-point (singleton) clusters and recursively merges two or more most appropriate clusters. A **divisive clustering** starts with one cluster of all data points and recursively splits the most appropriate cluster. The process continues until a

stopping criterion (frequently, the requested number *k* of clusters) is achieved.

Advantages of Hierarchical Method

- Embedded flexibility regarding the level of granularity.
- Ease of handling of any forms of similarity or distance.
- Consequently, applicability to any attributes types.

Disadvantages of Hierarchical Method

- Vagueness of termination criteria.
- The fact that most clustering algorithms do not revisit once constructed intermediate clusters with the purpose of improvement.

Partitioning Relocation Clustering

Partitioning Relocation Clustering divides data into several subsets. Because checking all possible subset systems is computationally infeasible, certain greedy heuristics are used in the form of *iterative optimization***.** Specifically, this means different *relocation* schemes that iteratively reassign points between the *k* clusters. Unlike traditional hierarchical methods, in which clusters are not revisited after being constructed, relocation algorithms gradually improve clusters. With appropriate data, this results in high quality clusters.

One approach to data partitioning is to take a **conceptual** point of view that identifies the cluster with a certain model whose unknown parameters have to be found. More specifically, **probabilistic models** assume that the data comes from a mixture of several populations whose distributions and priors we want to find. One clear advantage of probabilistic methods is the interpretability of the constructed clusters. Having concise cluster representation also allows inexpensive computation of intra-clusters measures of fit that give rise to a global *objective function.* Another approach starts with the definition of **objective function** depending on a partition. In iterative improvements such pair-wise computations would be too expensive. Using unique cluster representatives resolves the problem: now computation of objective function becomes linear in *N* (and in a number of clusters *k << N).* Depending on how representatives are constructed, iterative optimization partitioning algorithms are subdivided into **k-medoids and k-means** methods. K-medoid is the most appropriate data point within a cluster that represents it. Representation by k-medoids has two advantages. First, it presents no limitations on attributes types, and, second, the choice of medoids is dictated by the location of a predominant fraction of points inside a cluster and, therefore, it is lesser sensitive to the presence of outliers. In k-means case a cluster is represented by its centroid, which is a mean (usually weighted average) of points within a cluster. This works conveniently only with numerical attributes and can be negatively affected by a single outlier. On the other hand, centroids have the advantage of clear geometric and statistical meaning.

K-means Clustering Algorithm

An important facet in clustering is the similarity function that is used. When the data is numeric, a similarity function based on distance is typically used. For example, the Euclidean distance can be used to measure similarity. Consider two n-dimensional data points (records) r_j and r_k. We can consider the value for the i^{th} dimensions as r_{ji} and r_{ki} for the two records. The Euclidean distance between two points r_j and r_k in n-dimensional space is calculated as:

$$\text{Distance}(r_j, r_k) = (|r_{j1} - r_{k1}|^2 + |r_{j1} - r_{k1}|^2 + |r_{j1} - r_{k1}|^2)^{1/2}$$

The smaller the distance between two points, the greater is the similarity as we think of them. A classic clustering algorithm is the K-means algorithm.

Input: A database D, of m records, $r_1 \ldots r_m$ and a desired number of clusters k.

Output: Set of k clusters that minimize the squad error correction.

Begin: randomly choose k records as the centroids of k clusters. ;

repeat

assign each record r_i, to a cluster such that the distance between r_i and the cluster centroid (mean) is smallest among the k clusters.

recalculate the centroid (mean) for each cluster based on the records assigned to the cluster;

until no change;

End;

The algorithm begins by randomly choosing k records to represent the centroids (means), $m_1,\ldots,m_k$ of the clusters $C_1,\ldots,C_k$. All the records are placed in a given cluster based on the distance mi and record r_j is the smallest among all cluster means, then record r_j is placed in cluster C_i. Once all records have been initially placed in a cluster, the mean for each cluster is recomputed. Then the process repeats, by examining each record examining each record again and placing it in the cluster whose mean is

closest. Several iterations may be needed, but the algorithm will converge, although it may terminate at a local optimum. The terminating condition is usually the squad-error criterion. For clusters $C_1,\ldots,C_k$ with means $m_1,\ldots,m_k$, the error is defined as:

$$\text{Error} = \sum_{i=1}^{k} \Big(\sum_{\forall r_j \in C_i} \text{Distance}\,(r_j , m_i)^2 \Big)$$

Applications of Data Mining

Data mining technologies can be applied to a large variety of decision-making contexts in business. In particular, areas of significant payoffs are expected to include the following:

- **Marketing:** Applications include analysis of consumer behavior based on buying patterns; determination of marketing strategies including advertising, store location, and targeted mailing; segmentation of customers, stores, or products; and design of catalogs, store layouts, and advertising campaigns.

- **Finance:** Applications include analysis of creditworthiness of clients, segmentation of account receivables, performance analysis of finance investments like stocks, bonds, and mutual funds; evaluation of financing options; and fraud detection.

- **Manufacturing:** Applications involve optimization of resources like machines, man-power, and materials; optimal design for manufacturing process, shop-floor layouts, and products design, such as for automobiles based on customer requirements.

- **Health Care:** Applications include discovering patterns in radiological images, analysis of micro-array (gene-clip) experimental data to relate to diseases, analyzing side effects of drugs, and effectiveness of certain treatments; optimization of processes within a hospital, relating patient wellness data with doctor qualifications.

Future Research on Data Mining

There have been many data mining systems developed in recent years, all this trend of research and development on data mining is expected to be flourishing because the huge amounts of data have been collected in databases and the necessity of understanding and making good use of such data in decision making has served as the driving force in data mining.

The diversity of data, data mining tasks, and data mining approaches poses many challenging research issues on data mining. The design of data mining languages, the development of efficient and effective data mining methods and systems, the construction of interactive and integrated data mining environment, and the application of data mining techniques at solving large application problems are the important tasks for data mining researchers and data mining system and application developers.

Moreover, with the fast computerization of the society, the social impact of data mining should lot be under-estimated. When large amount of interrelated data are effectively analyzed from different perspectives, it can pose threats to the goal of protecting data security and guarding against the invasion of privacy. It is a challenging task to develop effective techniques for preventing the disclosure of sensitive information in data mining, especially as the use of data mining systems is rapidly increasing in domains ranging from business analysis, customer analysis to medicine and government.

In conclusion, we have seen that to understand and enhance the data warehousing and data mining process we have relied upon tools traditionally belonging to both statistics and computer science. As statistician William Shannon (1999) wrote:

I think there is a challenge for statisticians to start learning machine learning and computer science, and machine learners to start learning statistics. These two fields rightly fall under the broad umbrella of "data analysis."

Acknowledgement

We would like to express deep sense of gratitude to Mrs. Saumya Hegde, Lecturer, Computer Science and Engineering Department, NITK Surathkal for her valuable suggestions, guidance and encouragement for completion of this paper.

References

- Fundamentals of Database Systems (Fourth Edition) by Remez Elmasri, *University of Texas at Arlington* and Shamkant B. Navathe*, Georgia Institute of Technology*.

- An Introduction to DATABASE SYSTEMS (Seventh Edition) by C.J. Date.

- Building the Data Warehouse, John Wiley 1992 by W.H Inmon

- An overview of Data warehousing and OLAP Technology by Surajit Chaudhri, *Microsoft Research, Redmond* and

Umeshwar Dayal, *Hewlett-Packard Labs, Palo Alto.*

- Integration of Data Mining and Data Warehousing Technology by, Jiawei Han, *Database Systems Research Laboratory, School of Computing Science Simon Fraser University, Canada.*

- Survey of Clustering Data Mining Techniques by Pavel Berkhin, *Accrue Software, Inc.*

- Comparing Algorithms and Clustering Data: Components of the Data Mining Process by Glenn A. Growe, Department of Computer Science and Information Systems Grand Valley State University, Mackinac Hall Allendale, Michigan 49401.

- An Introduction to Data Mining by Kurt Thearling.

- Using AutoMed Metadata in Data Warehousing Environments by Hao Fan, Alexandra Poulovassilis, *School of Computer Science and Information Systems Birkbeck College, University of London.*

- Comparing Comparisons: Document Clustering Evaluation Using Two Manual Classifications by Magnus Rosell and Viggo Kann KTH Nada SE-100 44 Stockholm Sweden and Jan-Eric Litton MEB, Karolinska Institute SE-171 77 Stockholm Sweden.

- Defining Data Warehousing - What is it and who needs it? by Silvon Software, Inc.

- Research Problems in Data warehousing by Widom J., *Proc.4th Intl. CIKM Conf., 1995.*

- Research Issues in Data mining by M.C. Wu

- www.sciencedirect.com

- http://pwp.starnetinc.com/larryg/articles.html

- http://support.sas.com/rnd/warehousing/index.html

- http://www.olapcouncil.org

- http://www.statsoft.com/textbook/stdatmin.html

- http://www.thearling.com/index.htm