Data Warehousing And Data Mining

***Preeti Chaudhary**
****Ruchi Tripathi**

Abstract

This paper describes a new approach to fast multimedia information retrieval with data mining and data ware housing techniques. To tackle the key issues such as multimedia data indexing, similarity measures, search methods and query processing in retrieval for large multimedia data archives, we extend the concepts of conventional data warehouse and multimedia data warehouse for effective data representation and storage.
In this study the technological advances are making this vision a reality for many organizations. Here, we would be discussing about the benefits that and organization will get through the use of data mining. We will be discussing about the various stages about the predictive data mining like, the initial exploration, model building or pattern identification with validation/verification, and deployment. It discusses about the various strategic applications of data mining and data warehousing. In addition, we propose a fuzzy neural network to provide automatic and autonomous classification for the retrieval outputs by integrating fuzzy logic technology and the Back Propagation Feed Forward (BPFF) neural network. A series of case studies are reported to demonstrate the feasibility of the proposed method.

Key Words: Data mining, Data warehousing, Strategic applications, Predictive data mining, Query processing, Fuzzy neural networks.

1. Introduction: A data warehouse is a centralized database that captures information from various parts of an organization's business processes. This information can later be analyzed to determine predictive relationships through the use of data mining techniques.Generally; data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful 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. Data warehousing is combining data from multiple and usually varied sources into one comprehensive and easily manipulated database. Common accessing systems of data warehousing include queries, analysis and reporting. Because data warehousing creates one database in the end, the number of sources can be anything you want it to be, provided that the system can handle the volume, of course. The final result, however, is homogeneous data, which can be more easily manipulated. Since it is not reasonable to expect that a user will know the physical locations and/or DB identities and logon procedures, of all of the data sources that might be relevant for his/her

*Lecturer (Computer Science and Engineering Deptt) Kanpur Institute of Technology
**Lecturer (Business Administration Deptt) Kanpur Institute of Technology

information request, some form of front-end system, consisting of an interface, search engine, and integrated database system, needs to be developed to provide access to the potentially multiple, relevant DBSs. Since it is also unreasonable to require that the user access relevant data one system at a time using potentially varying local system query languages, there is a need to develop a common user query language and let the underlying query processor do the necessary query translations. Thus, a desirable MMIRS (and MDBS) would offer a single interface and query language to the data in any number of multimedia database systems and then integrate and rank the results from the user search query. In traditional distributed database literature for structured DBs, it is common to distinguish between types of distributed system architectures based on the degree to which the component schemas are/can be integrated.

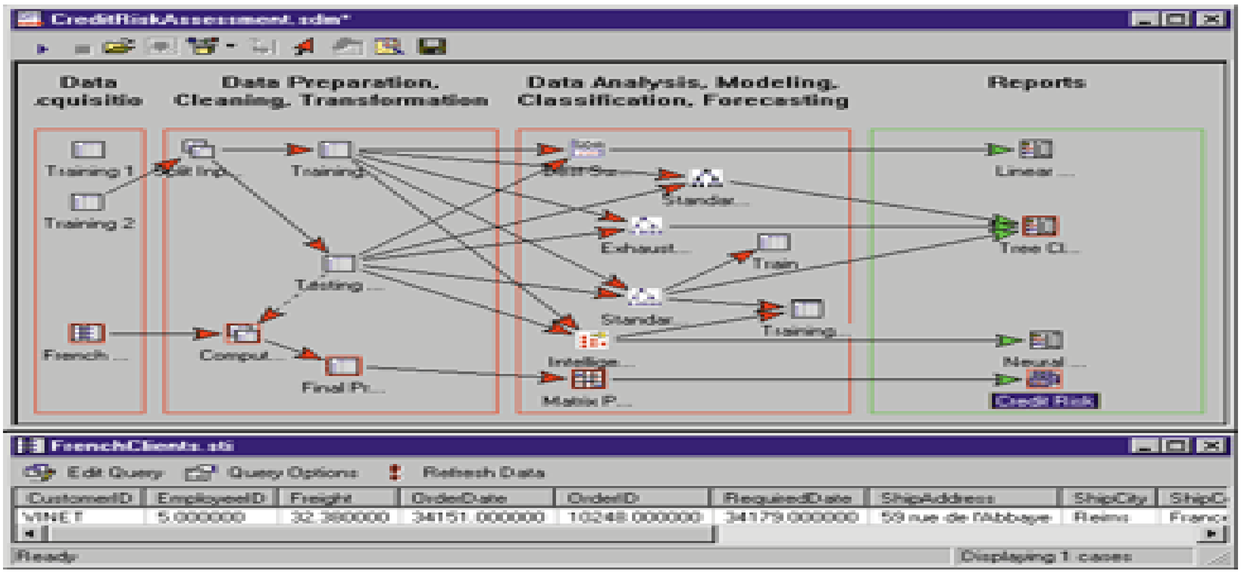
2. Component Schemas of Data Warehousing

- **Heterogeneous systems:** Heterogeneous multi-DB systems are predominately *loosely coupled* in the sense that they are 'constructed' as an integration of existing heterogeneous systems, each of which has been independently designed and implemented and is in use for a local application set. Independent design practically ensures that there will be semantic heterogeneity. A single *global* schema may exist, alternatively there may be a set of *federated schemas*. In either case, the integration schema is constructed through the union of the schemas for the participating databases. A synonym table and a thesaurus may be constructed to support single query access to the multiple component databases. The objective of the global or federated schema is to hide the diversity of structure, location and naming conventions used in the component schemas. Examples include the database set resulting from the merger of an organization or a *"union catalog"* digital library system, in which the union catalog functions as the global schema and participating libraries maintain existing on-line systems (OPACs). Another proposal is to use an integration of the metadata for participating media databases as the global access point. This system supports both visual and text queries to Web documents using a combination of image and text-based information retrieval with distributed (relational) database management. Using a similar approach and If we consider a Web-site to be a database, then search engines also support multi-database location. In this approach, search engine *crawlers* retrieve and index Web pages. The resulting indexes function as a global access point (or portal) to websites containing terms matching the query search terms. One can argue if the source data accessed by search engines really constitute a set of databases, or if the system 'simply' consists of a huge set of disparate Web pages for which the search engine's *crawler* has made a *term location* index that facilitates a keyword based query processor for Web page/site location.

- **Interoperable systems**: *Disjoint or language based* systems are very loosely connected. They have no or at best a very primitive, locally stored 'global' schema that defines the location and access paths to cooperating database systems. These systems have an extended query language processor that can access the local DB schema/metadata of cooperating systems and use *domain ontologies* to map a user query to relevant databases and documents within these. An alternative to installing a cooperative query processor at each location is to use agents for query interpretation and date retrieval. The **W3C** has also taken a language-based approach to data integration in their development of tools for the semantic web, which include XML, DTD, RDF schemas, and OWL for specifying Web ontologies. The latter are the key to 'understanding' Web data and for its integration. Much of W3C's focus in this area has been on developing *tag-based* tools to facilitate exchange of Web based data, 'on line' or extracted from underlying databases, from one application to another - so called "peer-to-peer" communication. The strategy used is to 'package' each data element within a tag set defined by XML and its DTD or RDF schema that is accessible to both the sender and receiver. The primary application area has been that of e-commerce. Less focus has been (to date) placed on access to multiple heterogeneous underlying databases.

- **Synonym Identification and Resolution Strategies :**The central problem in working with or creating multi-database systems is that of identifying and resolving the ***semantic heterogeneity*** that exists between the component databases). Semantic heterogeneity exists whenever databases are designed independently, over time, by different design teams and/or in different organizations. It is represented, in structured (relational) databases by differing attribute names and structures used to model the same data and/or concepts in different systems. It is formalized, and to some degree recognizable, in the individual data models and schemas used to implement the set of component databases.
  Semantic heterogeneity also exists between collections of semi-structured and unstructured data, such as between different XML document collections and image or text metadata. Thus, *semantic heterogeneity* exists whenever there is more than one way to structure a data collection and is a problem whenever one wants integrated access to multiple data collections.

3. Data Mining: Data Mining is an analytic process designed to explore data (usually large amounts of data - typically business or market related) in search of consistent patterns and/or systematic relationships between variables, and then to validate the findings by applying the detected patterns to new subsets of data. The ultimate goal of data mining is prediction - and predictive data mining is the most common type of data mining and one that has the most direct business applications. The process of data mining consists of three stages: (1) **the initial exploration**, (2) model building or pattern identification with **validation/verification**, and (3) **deployment** (i.e., the application of the model to new data in order to generate predictions).



CustomerID	EmployeeID	Freight	OrderDate	OrderID	RequiredDate	ShipAddress	ShipCity	ShipC
VINET	5.000000	32.380000	34151.000000	10248.000000	34179.000000	59 rue de l'Abbaye	Reims	Franc

Stage 1: Exploration. This stage usually starts with data preparation which may involve cleaning data, data transformations, selecting subsets of records and - in case of data sets with large numbers of variables ("fields") - performing some preliminary **feature selection** operations

to bring the number of variables to a manageable range (depending on the statistical methods which are being considered). Then, depending on the nature of the analytic problem, this first stage of the process of data mining may involve anywhere between a simple choice of straightforward predictors for a regression model, to elaborate exploratory analyses using a wide variety of graphical and statistical methods in order to identify the most relevant variables and determine the complexity and/or the general nature of models that can be taken into account in the next stage.

Stage 2: Model building and validation. This stage involves considering various models and choosing the best one based on their predictive performance (i.e., explaining the variability in question and producing stable results across samples). This may sound like a simple operation, but in fact, it sometimes involves a very elaborate process. There are a variety of techniques developed to achieve that goal - many of which are based on so-called "competitive evaluation of models," that is, applying different models to the same data set and then comparing their performance to choose the best. These techniques - which are often considered the core of **predictive data mining** - include: **Bagging** (Voting, Averaging), **Boosting**, **Stacking (Stacked Generalizations)**, and **Meta-Learning**.

Stage 3: Deployment. That final stage involves using the model selected as best in the previous stage and applying it to new data in order to generate predictions or estimates of the expected outcome.

The concept of *Data Mining* is becoming increasingly popular as a business information management tool where it is expected to reveal knowledge structures that can guide decisions in conditions of limited certainty. Recently, there has been increased interest in developing new analytic techniques specifically designed to address the issues relevant to business *Data Mining* (e.g., **Classification Trees**), but Data Mining is still based on the conceptual principles of statistics including the traditional **Exploratory Data Analysis (EDA)** and modeling and it shares with them both some components of its general approaches and specific techniques.

However, an important general difference in the focus and purpose between Data Mining and the traditional Exploratory Data Analysis (EDA) is that Data Mining is more oriented towards applications than the basic nature of the underlying phenomena. In other words, Data Mining is relatively less concerned with identifying the specific relations between the involved variables. For example, uncovering the nature of the underlying functions or the specific types of interactive, multivariate dependencies between variables are not the main goal of Data Mining. Instead, the focus is on producing a solution that can generate useful predictions. Therefore, Data Mining accepts among others a "black box" approach to data exploration or knowledge discovery and uses not only the traditional Exploratory Data Analysis (EDA) techniques, but also such techniques as **Neural Networks** which can generate valid predictions but are not capable of identifying the specific nature of the interrelations between the variables on which the implications are dependent.

4. Crucial Concepts in Data Mining:

Bagging (Voting, Averaging): The concept of bagging (voting for classification, averaging for regression-type problems with continuous dependent variables of interest) applies to the area of **predictive data mining**, to combine the predicted classifications (prediction) from multiple models, or from the same type of model for different learning data. It is also used to address the inherent instability of results when applying complex models to relatively small data sets. Suppose your data mining task is to build a model for predictive classification, and the dataset from which to train the model (learning data set, which contains observed classifications) is relatively small. You could repeatedly sub-sample (with replacement) from the dataset, and apply, for example, a tree classifier (e.g., **C&RT** and **CHAID**) to the successive samples. In practice, very different trees will often be grown for the different samples, illustrating the instability of models often evident with small data sets. One method of deriving a single prediction (for new observations) is to use all trees found in the different samples, and to apply some simple voting: The final classification is the one most often predicted by the different trees. Note that some weighted combination of predictions (weighted vote, weighted average) is also possible, and commonly used. A sophisticated (**machine learning**) algorithm for generating weights for weighted prediction or voting is the **Boosting procedure**.

Boosting: The concept of boosting applies to the area of **predictive data mining**, to generate multiple models or classifiers (for prediction or classification), and to derive weights to combine the predictions from those models into a single prediction or predicted classification A simple algorithm for boosting works like this: Start by applying some method (e.g., a tree classifier such as **C&RT** or **CHAID**) to the learning data, where each observation is assigned an equal weight. Compute the predicted classifications, and apply weights to the observations in the learning sample that are inversely proportional to the accuracy of the classification. In other words, assign greater weight to those observations that were difficult to classify (where the misclassification rate was high), and lower weights to those that were easy to classify (where the misclassification rate was low). In the context of C&RT for example, different misclassification costs (for the different classes) can be applied, inversely proportional to the accuracy of prediction in each class. Then apply the classifier again to the weighted data (or with different misclassification costs), and continue with the next iteration (application of the analysis method for classification to the re-weighted data).
Boosting will generate a sequence of classifiers, where each consecutive classifier in the sequence is an "expert" in classifying observations that were not well classified by those preceding it. During **deployment** (for prediction or classification of new cases).

Data Preparation (in Data Mining): Data preparation and cleaning is an often neglected but extremely important step in the data mining process. The old saying "garbage-in-garbage-out" is particularly applicable to the typical data mining projects where large data sets collected via some automatic methods (e.g., via the Web) serve as the input into the analyses. Often, the method by which the data where gathered was not tightly controlled, and so the data may contain out-of-range values (e.g., Income: -100), impossible data combinations (e.g., Gender: Male, Pregnant: Yes), and the like. Analyzing data that has not been carefully screened for such problems can produce highly misleading results, in particular in predictive data mining.

Data Reduction (for Data Mining): The term Data Reduction in the context of data mining is usually applied to projects where the goal is to aggregate or amalgamate the information contained in large datasets into manageable (smaller) information nuggets. Data reduction methods can include simple tabulation, aggregation (computing descriptive statistics) or more sophisticated techniques like **clustering**, **principal components analysis**, etc.

Predictive Data Mining combines database analysis with multivariate statistics and artificial intelligence. In recent years, predictive data mining has become an essential tool for strategic decision making among mid-size to large corporations. It has been proven effective in predicting future customer behavior, classifying customer segments and forecasting events.
Right now, there are huge databases and powerful technologies working together to crunch numbers about your lifestyle and lifestyles of millions of other Americans. They know the value of your home, the type of car you drive, the ages of your children, your credit rating and more. This data is being mathematically processed to determine if you are the best target for the latest gadget to hit the market.

While this sounds like something from a George Orwell novel, it describes the **predictive modeling** power behind currently available modern data mining technology. While data mining conducted at this magnitude is limited to certain government agencies, the price of this technology has dropped substantially due to new mathematical discoveries, lower technology costs and improved processing power.

5. Autonomous Classification for The Retrieval Outputs

Fuzzy Logic and Neural Networks: Fuzzy logic is reasoning with uncertainty. That is, instead of a two valued logic (true or false), there are multiple values (true, false, maybe). Fuzzy logic has been used in database systems to retrieve data with imprecise or missing values. In this case, the membership of records in the query result set is fuzzy.

Fuzzy Logic is a problem-solving control system methodology that lends itself to implementation in systems ranging from simple, small, embedded micro-controllers to large, networked, multi-channel PC or workstation-based data acquisition and control systems. It can be implemented in hardware, software, or a combination of both. FL provides a simple way to arrive at a definite conclusion based upon vague, ambiguous, imprecise, noisy, or missing input information. FL's approach to control problems mimics how a person would make decisions, only much faster. The combination of fuzzy logic and neural network has resulted in an extremely powerful computational model known as fuzzy neural network. Representing a linguistic value as a fuzzy set has enabled the system to deal successfully with many expert problems. Neural heuristics further provide the fuzzy network with the capability of learning by self-adaption and self-organization.

The Feed-Forward Neural Network Model: If we consider the human brain to be the 'ultimate' neural network, then ideally we would like to build a device which imitates the brain's functions. However, because of limits in our technology, we must settle for a much simpler design. The obvious approach is to design a small electronic device which has a transfer function similar to a biological neuron, and then connect each neuron to many other neurons, using RLC networks to imitate the dendrites, axons, and synapses. This type of electronic model is still rather complex to implement, and we may have difficulty 'teaching' the network to do anything useful. Further constraints are needed to make the design more manageable. First, we change the connectivity between the neurons so that they are in distinct layers, such that each neruon in one layer is connected to every neuron in the next layer. Further, we define that signals flow only in one direction across the network, and we simplify the neuron and synapse design to behave as analog comparators being driven by the other neurons through simple resistors. We now have a feed-forward neural network model that may actually be practical to build and use.

Referring to figures 1 and 2, the network functions as follows: Each neuron receives a signal from the neurons in the previous layer, and each of those signals is multiplied by a separate weight value. The weighted inputs are summed, and passed through a limiting function which scales the output to a fixed range of values. The output of the limiter is then broadcast to all of the neurons in the next layer. So, to use the network to solve a problem, we apply the input values to the inputs of the first layer, allow the signals to propagate through the network, and read the output values.

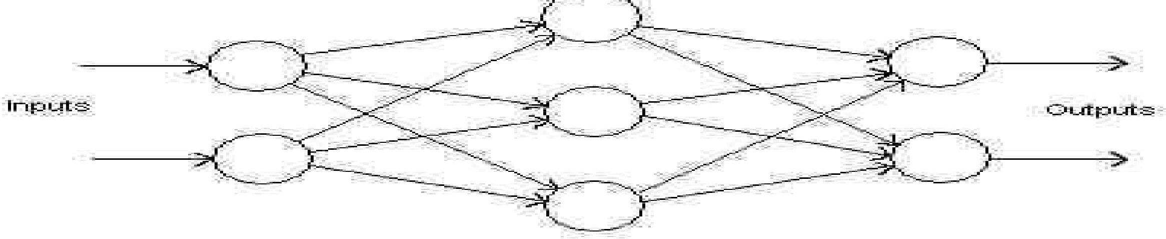


Figure 1. A Generalized Network. Stimulation is applied to the inputs of the first layer, and signals propagate through the middle (hidden) layer(s) to the output layer. Each link between neurons has a unique weighting value.

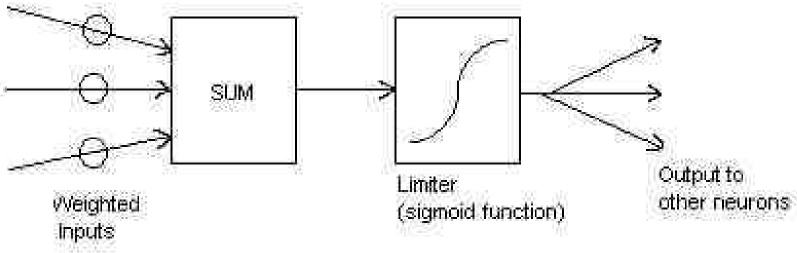


Figure 2. The Structure of a Neuron. Inputs from one or more previous neurons are individually weighted, then summed. The result is non-linearly scaled between 0 and +1, and the output value is passed on to the neurons in the next layer.

Since the real uniqueness or 'intelligence' of the network exists in the values of the weights between neurons, we need a method of adjusting the weights to solve a particular problem. For this type of network, the most common learning algorithm is called Back Propagation (BP). A BP network learns by example, that is, we must provide a learning set that consists of some input examples and the known-correct output for each case. So, we use these input-output examples to show the network what type of behavior is expected, and the BP algorithm allows the network to adapt.

The BP learning process works in small iterative steps: one of the example cases is applied to the network, and the network produces some output based on the current state of it's synaptic weights (initially, the output will be random). This output is compared to the known-good output, and a mean-squared error signal is calculated. The error value is then propagated backwards through the network, and small changes are made to the weights in each layer. The weight changes are calculated to reduce the error signal for the case in question. The whole process is repeated for each of the example cases, then back to the first case again, and so on. The cycle is repeated until the overall error value drops below some pre-determined threshold. At this point we say that the network has learned the problem "well enough" - the network will never exactly learn the ideal function, but rather it will asymptotically approach the ideal function.

6. Data Warehousing and Data Mining In The Market: Data warehousing is commonly used by companies to analyze trends over time. In other words, companies may very well use data warehousing to view day-to-day operations, but its primary function is facilitating strategic planning resulting from long-term data overviews. From such overviews, business models, forecasts, and other reports and projections can be made. Routinely, because the data stored in data warehouses is intended to provide more overview-like reporting, the data is read-only. If you want to update the data stored via data warehousing, you'll need to build a new query when you're done.

Data warehousing is typically used by larger companies analyzing larger sets of data for enterprise purposes. Smaller companies wishing to analyze just one subject, for example, usually access data marts, which are much more specific and targeted in their storage and reporting. Data warehousing often includes smaller amounts of data grouped into data marts. In this way, a larger company might have at its disposal both data warehousing and data marts, allowing users to choose the source and functionality depending on current needs.

7. Marketing Predictions: Producing **accurate forecasts** is an important part of measuring your marketing strategy. Inaccurate forecasts lead to increased inventory costs, under or over production, missed targets, improperly allocated resources and many other problems. While tools like Microsoft Excel provide some forecasting tools, the accuracy of these tools are significantly reduced when non-linear relationships or missing data are present, which is often the case when analyzing marketing data. In many cases, neural networks can provide superior forecasting accuracy.

8. **Market Segmentation:** When neural networks are setup appropriately, they can accurately identify people who will be most receptive to a product, promotion or advertising campaign. Some of the most frequent methods of segmentation with neural networks combine metrics such as recency of purchase, frequency of purchases and amount spent. Other factors include age, sex, income, location, education level, occupation and household status. Today, neural networks are a primary method for highly predictive marketing segmentation.

9. Prediction and Classification: Neural networks are a proven technology for solving complex classification problems. Credit companies often deploy neural networks to spot fraudulent credit card activity and identity theft. Other companies deploy neural networks to identify defecting customers in order to maximize their customer retention.
The Marketing Analysts are experienced in deploying neural networks to discover marketing opportunities, segment customers and enable you to discover more complex relationships in your data. With our technology, we can develop more accurate and effective predictive models for better decision-making. Contact us today for more information.

10. Applications of Data Warehousing And Data Mining

Sales/Marketing:

Identify buying patterns from customers
Find associations among customer demographic characteristics
Predict response to mailing campaigns
Market basket analysis

Banking:

Credit card fraudulent detection
Identify ‘loyal’ customers
Predict customers likely to change their credit card affiliation
Determine credit card spending by customer groups
Find hidden correlation’s between different financial indicators
Identify stock trading rules from historical market data

Insurance and Health Care:

Claims analysis i.e., which medical procedures are claimed together
Predict which customers will buy new policies
Identify behaviour patterns of risky customers
Identify fraudulent behaviour

Transportation:

Determine the distribution schedules among outlets
Analyze loading patterns

Medicine:

Characterize patient behaviour to predict office visits
Identify successful medical therapies for different illnesses

11. Applications in Neural Networks: Neural networks are applicable in virtually every situation in which a relationship between the predictor variables (independents, inputs) and predicted variables (dependents, outputs) exists, even when that relationship is very complex and not easy to articulate in the usual terms of "correlations" or "differences between groups." A few representative examples of problems to which neural network analysis has been applied successfully are:

- **Detection of medical phenomena.** A variety of health-related indices (e.g., a combination of heart rate, levels of various substances in the blood, respiration rate) can be monitored. The onset of a particular medical condition could be associated with a very complex (e.g., nonlinear and interactive) combination of changes on a subset of the variables being monitored. Neural networks have been used to recognize this predictive pattern so that the appropriate treatment can be prescribed.
- **Stock market prediction.** Fluctuations of stock prices and stock indices are another example of a complex, multidimensional, but in some circumstances at least partially-deterministic phenomenon. Neural networks are being used by many technical analysts to make predictions about stock prices based upon a large number of factors such as past performance of other stocks and various economic indicators.
- **Credit assignment.** A variety of pieces of information are usually known about an applicant for a loan. For instance, the applicant's age, education, occupation, and many other facts may be available. After training a neural network on historical data, neural network analysis can identify the most relevant characteristics and use those to classify applicants as good or bad credit risks.
- **Monitoring the condition of machinery.** Neural networks can be instrumental in cutting costs by bringing additional expertise to scheduling the preventive maintenance of machines. A neural network can be trained to distinguish between the sounds a machine makes when it is running normally ("false alarms") versus when it is on the verge of a problem. After this training period, the expertise of the network can be used to warn a technician of an upcoming breakdown, before it occurs and causes costly unforeseen "downtime."
- **Engine management.** Neural networks have been used to analyze the input of sensors from an engine. The neural network controls the various parameters within which the engine functions, in order to achieve a particular goal, such as minimizing fuel consumption.

12. Conclusion: From a database perspective, where emphasis is placed on basic data mining concepts and techniques for uncovering interesting data patterns oriented towards the developments of multimedia data indexing, conventional data warehouse and multimedia data warehouse for effective representation. Various Strategic applications of data mining and data warehousing used in companies to process the volumes of available data within a company in a meaningful and reliable way, any company considering implementing a data warehouse or data mart will have to anticipate a growing monster that will require more IT/IS staff than they currently employ and will be marginally reliable in reporting "nuggets of market-savvy truth".
The combination of fuzzy logic and neural networks has results in an extremely powerful computational model i.e. fuzzy neural network. Neural heuristics further provide the fuzzy network and back propagation feed forward neural network with the capability of learning by self-adaptation and self-organization. As a result, many corporations are embracing predictive data mining to segment customers, predict customer behavior and make future projections based on historical data. Thus Data mining and Data Warehousing are the important ingredients of any emerging organization.

13. References

1. Bezdek, J.(ed.),(1992) “Fuzzy Logic and neural networks”.
2. Bishop, C. (1995), “ *Neural Networks for Pattern Recognition”*. Oxford: University Press. Extremely well-written, up-to-date. Requires a good mathematical background, but rewards careful reading, putting neural networks firmly into a statistical context.

3. Carling, A. (1992), "*Introducing Neural Networks*". Wilmslow, UK: Sigma Press. A relatively gentle introduction. Starting to show its age a little, but still a good starting point.
4. Dunham. Margaret H. and Sridhar, S., (1991)Data Mining An Introductory and Advanced Topic
5. Europe Gets into Fuzzy Logic" (Electronics Engineering Times, Nov. 11, 1991).
6. Fausett, L. (1994), "*Fundamentals of Neural Networks*". New York: Prentice Hall. A well-written book, with very detailed worked examples to explain how the algorithms function.
7. "Fuzzy Sets and Applications: Selected Papers by L.A. Zadeh", ed. R.R. Yager et al. (John Wiley, New York, 1987).
8. Haykin, S. (1994), "*Neural Networks: A Comprehensive Foundation*". New York: Macmillan Publishing. A comprehensive book, with an engineering perspective. Requires a good mathematical background, and contains a great deal of background theory.
9. Jiawei Han and Micheline Kamber. Data Mining
10. LiMin Fu. Neural Networks in Computer Intelligence.
11. Patterson, D. (1996), "*Artificial Neural Networks*". Singapore: Prentice Hall. Good wide-ranging coverage of topics, although less detailed than some other books.
12. "U.S. Loses Focus on Fuzzy Logic" (Machine Design, June 21, 1990).
13. "Why the Japanese are Going in for this 'Fuzzy Logic'" by Emily T. Smith (Business Week, Feb. 20, 1993, pp. 39).