Design and Implementation of an Intelligent Information Infrastructure

Henry C.W. Lau, Andrew Ning, and Peggy Fung

Department of Industrial and Systems Engineering,
The Hong Kong Polytechnic University, Hunghom, Hong Kong

Abstract. The lack of seamless data interchange and efficient data analysis hinders the formation of an effective information infrastructure that serves as the platform for data interchange across heterogeneous database systems. Information infrastructure has become an emerging platform to enable business partners, customers and employees of enterprises to access and interchange corporate data from dispersed locations all over the world. In general, information infrastructure is a browser-based gateway that allows users to gather, share, and disseminate data through Internet easily. This paper proposes the design and implementation of an information infrastructure embracing the emerging eXtensible Markup Language (XML) open standard, together with an intelligent data mining technique combining neural networks and On-Line Analysis Process (OLAP) and rule-based reasoning approaches to support knowledge discovery. To validate the feasibility of this approach, an information infrastructure prototype is developed and tested in a company with description of this case example covered in this paper.

1 Background

Recently, data mining technology, which aims at the conversion of clusters of complex data into useful information, has been under active research [1,15,23,24]. Data mining, in general, identifies and characterizes interrelationships among multivariable dimensions without requiring human effort to formulate specific questions. In other words, data mining is concerned with discovering new, meaningful information, so that decision-makers can learn as much as they can from their valuable data assets. Data mining tools require different data formats in relational and multi-dimensional database systems. The shared data access interface of data mining tools will enable easier exchange of information among different sources. There are many commercial products for data mining and a typical example of such product is Microsoft SQL server which has incorporated the On-Line Analytical Processing (OLAP) technology that provides a service for accessing, viewing and analyzing on large volume of data with high flexibility and performance [23,27].

T.D. Gedeon and L.C.C. Fung (Eds.): AI 2003, LNAI 2903, pp. 1051-1062, 2003.

While OLAP is able to provide numerical and statistical analysis of data in an efficient and timely way, it lacks the predictive capability, such as the projection of possible outcomes based on the past history of events so as to decide on the action to be taken. In this respect, it seems necessary that a certain "ingredient" of intelligence element needs to be added to OLAP to enable the self-learning capability of the whole system.

Neural network is a technology that has typically been used for prediction, clustering, classification and alerting of abnormal pattern [7,26]. They create predictive networks by considering a "training set" of actual records. In theory, the formation of neural network is similar to the formation of neural pathways in the brain as a task is practiced. Also a neural network refines its network with each new input it considers. To predict a future scenario, the neural network technology is able to work with a training set of data from the past history of records. It will use the training set to build a network, based on which the neural network is able to predict future scenario by supplying a set of attributes. Like the OLAP technology, there are many commercial products that have incorporated neural network technology. The inclusion of computational intelligence knowledge into a data mining technology can significantly enhance the "machine learning" capability of the system and is undoubtedly an issue that is justified to be addressed.

In creating decision support functionality, a mechanism, which is able to combine and coordinate many sets of diversified data into a unified and consistent body of useful information, is required. In larger organizations, many different types of users with varied needs must utilize the same massive data warehouse to retrieve the right information for the right purpose. Whilst data warehouse is referred as a very large repository of historical data pertaining to an organization, data mining is more concerned with the collection, management and distribution of organized data in an effective way. The nature of a data warehouse includes integrated data, detailed and summarized data, historical data and metadata. Integrated data enable the data miner to easily and quickly look across vistas of data. Detailed data is important when the data miner wishes to examine data in its most detailed manner while historical data is essential because important information nuggets are hidden in this type of data. OLAP is an example of architectural extension of the data warehouse.

Since after the setup of a data warehouse, the attention is usually switched to the area of data mining, which aims to extract new and meaningful information. In other words, a pool of 'useful information' that has been stored in a company data warehouse becomes 'intelligent information', thereby allowing decision-makers to learn as much as they can from their valuable data assets. In this respect, neural network can be deployed to enhance the intelligence level of the OLAP application.

Neural network searches for hidden relationships, patterns, correlation and interdependencies in large databases that traditional information gathering methods (such as report creation and user querying) may have overlooked. The responsibility of the neural network is to provide the desire change of parameters based on what the network has been trained on. Intrinsically, a sufficient amount of data sample is a key factor in order to obtain accurate feedback from the trained network. As neural network is meant to learn relationships between data sets by simply having sample data represented to their input and output layers [9], the training of the network with input

and output layers mapped to relevant realistic values with the purpose to develop the correlation between these two groups of data will not, in principle, contradict the basic principle of neural network.

With a trained network available, it is possible that recommended action can be obtained with the purpose to rectify some hidden problems, should that occur at a later stage. Therefore, in the training process of the neural network, the nodes of the input layer of the neural network represent the data from the OLAP and those of the output layer represent the predictions and extrapolations. The data flow of the NOLAPS has been depicted in Fig. 1. It should be noted that the output information from the OLAP could be used to refine the OLAP data cube so as to continually update the database over time.

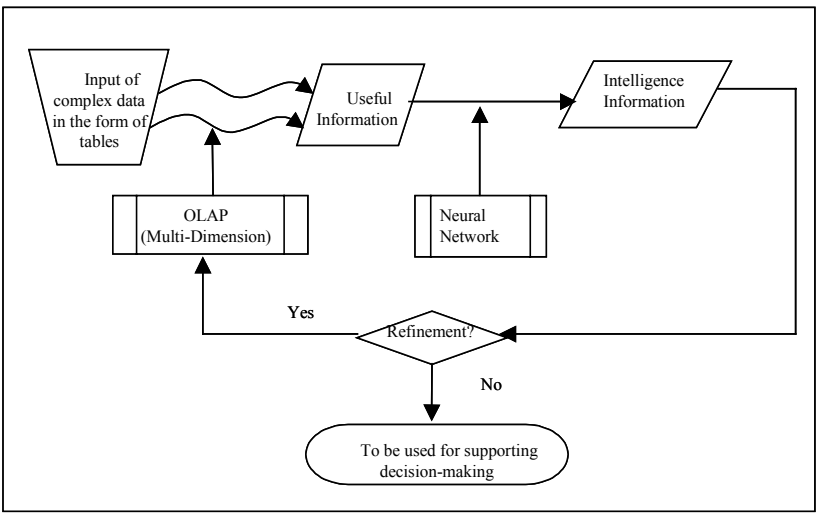

Fig. 1. Information flow of NOLAP

The data interchange within the NOLAPS encompasses three modules, namely OLAP module, Data Conversion (DC) module and Neural Network (NN) module (Fig. 2). The data repository, which aims to support efficient data interchange among the three modules, is essential for the coordination and updating of information from various sources. As for the OLAP module, it consists of descriptive data (dimensions) and quantitative value (measures), both of which generate the OLAP data cube by building up two elements, namely, fact table and dimension [6]. In the fact table, the required data and user-defined methods for analysis are specified clearly. In the descriptive data of OLAP, the different dimension levels are defined for further computational use on different views of OLAP data cube. Typical dimension includes location, company and time whereas typical measure includes price, sales and profit. With a multidimensional view of data, the OLAP module provides the foundation for analytical processing through flexible access to information. In particular, this distinct feature can be used to compute a complex query and analyze data on reports, thereby achieving the viewing of data in different dimensions in a more easy and efficient way.

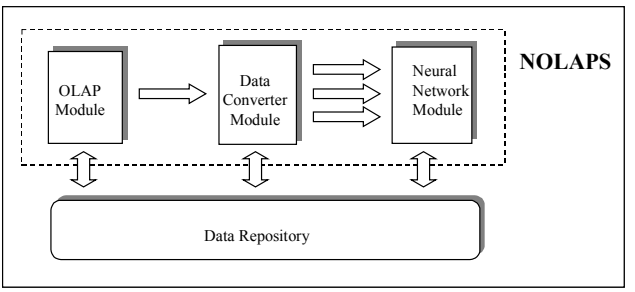

Fig. 2. Infrastructure of NOLAPS

2 Implementation of Neural Online Analytical Processing System

A prototype system has been developed, based on the framework of the NOLAPS. Pursuing the NOLAPS infrastructure that has been defined in the previous section, the OLAP module has generated a pool of useful data and accordingly, the NN module has created a reliably trained neural network. Next, 5 latest track records of a company has been gathered and listed as follows.

Performance Score Point (PSP) ranging from 1 (least point) to 7 (highest point) is used to assess the company as shown below.

Company A	Product quality PSP	Product cost PSP	Delivery schedule PSP
Latest record	3.5	6.5	6.6
2nd latest record	4.7	5.5	5.4
3rd latest record	5.0	5.1	4.8
4th latest record	5.6	4.1	4.4
5th latest record	4.0	4.0	3.0

After such information has been input, the NN module gives an assessment report back to user, thus supporting user to take action if deemed necessary. In the following table, "0" output from the NN node indicates a negative suggestion to the associated statement and "1" is the positive suggestion whereas "0.5" indicates that there is not enough data to justify a firm suggestion.

Company A	Output from NN module
Potentially competent	0.5
Suggested to be replaced	0
Service quality is compromised to meet the quoted price	1
Further assessment of company performance is required	1
Delivery time seems to be inconsistent due to certain company problems	1

Referring to Figures 3 & 4 and based on the NN output results as shown in the table, Company A seems to have problem in meeting the delivery schedules and it is suggested that further assessment regarding the company's performance is needed. Based on the suggestion of this assessment report, Company A has been approached in order to find out the reason behind the continual downgrade of performance in terms of product quality. After an organized investigation of the issue, it has been found that several of the senior staff of the product quality assurance group have left the company to start their own business. Because of this unexpected change, the company has suffered an unprecedented "brain-drain", resulting in the sudden decline of quality level of certain mainstream products.

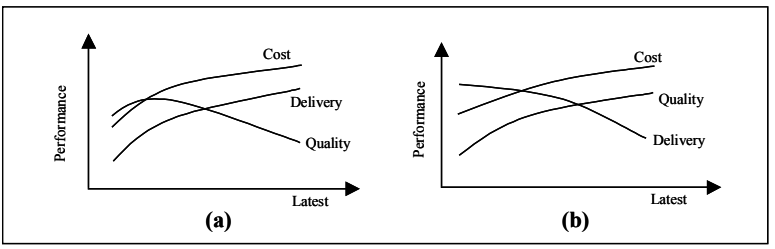

Fig. 3. Performance of a company based on past records

Because of the situation, Company A has been suggested to adopt some best practices related to quality assurance. In this case, the Total Quality Management (TQM) practice has been adopted and some necessary tools have also been acquired in order to implement such practice in the company. At this stage, it is still difficult to tell if the company could significantly reverse the downturn performance in terms of product quality. However, because of the signal generated from the NOLAPS, the problem of a business partner has been alerted and quick decision has been made with supporting assessment report, thus avoiding the loss of a trusted business partner, which can in turn weaken the overall performance of the VEN.

This case example indicates that the introduction of the neural network module to the OLAP module is able to significantly upgrade the decision support functionality. However, the results obtained, so far, is by no means perfect although it demonstrates that the suggested NOLAPS is basically viable and therefore it is justifiable to have further investigation along this line of research.

3 Neural Fuzzy Model

Fuzzy logic is a superset of conventional (Boolean) logic that has been extended to handle the concept of partial truth. According to Zadehi [29,30,31], rather than regarding fuzzy theory as a single theory, people should regard the process of "fuzzification" as a methodology to generalize any specific theory from a crisp (discrete) to a continuous (fuzzy) form. In particular, fuzzy logic has been deployed to replace the role of mathematical model with another that is built from a number of rules with fuzzy variables such as output temperature and fuzzy terms such as relatively high and reasonably low [2,3,12,22].

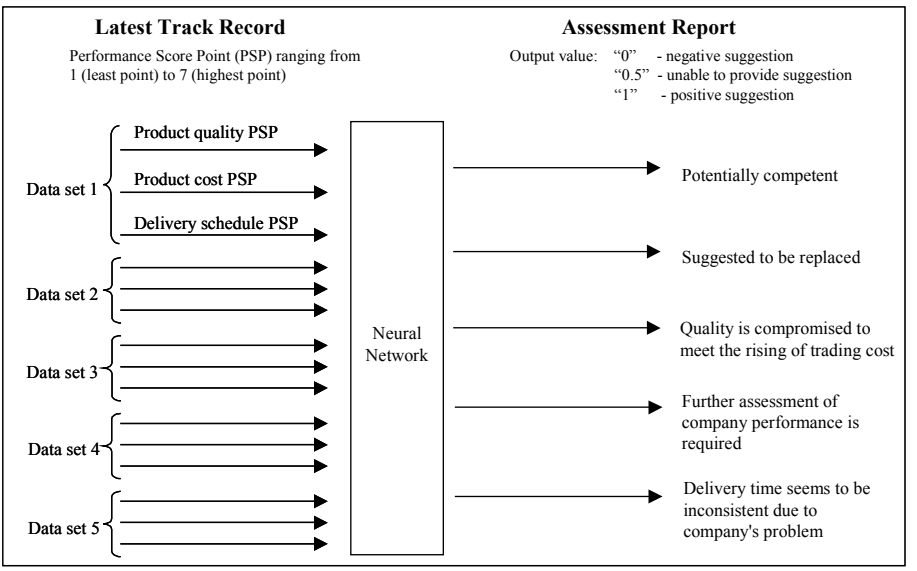

Fig. 4. Mapping of input and output nodes of Neural Network (NN) module of the NOLAPS

In an environment that there are multiple input and output parameters interacting with each others, many authors suggest that a fuzzy logic inference architecture can be employed to deal with the complex control issues [4,11,12,17,21,28]. Driankov *et al* [5] suggests the computational structure of a Fuzzy Knowledge base Controller (FKBC) to handle parameter-based control, which encompasses five computational steps including input scaling (normalization), fuzzification of inputs, inference or rule firing, defuzzification of outputs and output denormalization. It is not the intention of this paper to justify any detailed analysis, methodologies and techniques behind the fuzzy inference architectures covered in the above articles. However, most of the techniques described are very useful in understanding the design and operations of the fuzzy inference architecture that the development of a fuzzy expert system can be based on.

Whilst most of the fuzzy rules of these techniques are being set based on experience, past history and theoretical background, it is obvious that more can be done regarding the formulation of an algorithmic solution and the structure finding from existing data, which may lead to any generation of rules. In this respect, neural network has the ability to learn the relationship among input and output data sets through a training process, thus is able to "induce" output data if a new set of input data is made available. Although it can be said that neural network cannot do anything that cannot be done using traditional computing techniques, but they can handle some tasks that would otherwise be very difficult. For example, the rules generated by the neural network training process are more acceptable than those decided by expert knowledge that may not be independent and fair [31,32].

While fuzzy logic systems allow the use of linguistic terms representing data sets in the reasoning process, neural network is able to discover connections between data sets by simply having sample data represented to its input and output layers [8]. Neu-

ral network can be regarded as processing device, and it usually has some sort of "training" rule whereby the weights of connections are adjusted on the basis of presented patterns.
In order to enhance machine intelligence, an integrated neural-fuzzy model, which aims to make use of benefits generated by the synergy of neural network and fuzzy logic intelligence techniques. The main feature of the neural-fuzzy model is that it uses the trained neural network to generate 'If-Then' rules, which are then fuzzified prior to undergoing a fuzzy inferencing process, resulting in the generation of new process parameters for enhancing machine intelligence. The neural-fuzzy system described in this paper adopts inductive learning through the network, indicating that the system induces the information in its knowledge base by example. That induced information is then fuzzified and fed to the fuzzy logic system prior to fuzzy reasoning process. The significance of the neural-fuzzy model is that it attempts to formulate a technique for eliminating the knowledge acquisition bottleneck, thus enhancing the intelligent monitoring of a parameter-based control situation.

4 Demonstration of Neural-Fuzzy Model

The responsibility of the neural network model element is to provide the desire change of parameters based on what the network has been trained on. Intrinsically, a sufficient amount of data sample is a key factor in order to obtain accurate feedback from the trained network. In actual situations, recommended action about the required change of parameters to cope with the dimensional inconsistency is essential. In view of this situation, neural network can be regarded as a better option, if the dimensional values are mapped to the nodes of the input layer and heat transfer parameters are mapped to the output layer nodes, thus resulting in a control model that is the reverse of the heat transfer model. In the light of the fact that in an actual thermal system design, the required overall heat transfer is first determined from the system analysis. Then the rib geometry is chosen according to the nearest overall heat transfer performance determined from experimental investigations. Very often the difference between the designed overall heat transfer and the experimental performance data can be quite significant.

With a neural network, the correlation between the deviations of heat transfer parameters in response to the deviations of the occurring dimensional values can be trained based on a wide spectrum of actual sample data. As neural network is intended to learn relationships between data sets by simply having sample data represented to their input and output layers [9], the training of a network with input and output layers mapped to dimensional deviation values and heat transfer deviation values respectively with the purpose to develop the correlation between these two groups of data will not contradict the basic principle of neural network.

With a trained network available, it is possible that recommended action about the change of parameters can be obtained with the purpose to optimize the design of rib geometry, should that occur at a later stage. Therefore, in the training process of the neural network, the nodes of the input layer of the neural network represent the devia-

tion of the dimensional values and those of the output layer represent the deviation of the heat transfer parameters.

If there is dimensional inconsistency on the heat transfer model, the values at the nodes from the neural network (representing the parameter deviations) may provide some hints for possible dimensional correction. With the availability of this information, a fuzzy logic approach can then be employed to provide a modified set of recommended parameter change based on the original output values from the neural network. The motive for using fuzzy logic reasoning in this model is to take advantage of its ability to deal with imprecision terms which fit ideally in the parameter-based control situations where terms such as "rib spacing could be increased slightly" are used. Furthermore, the vagueness and uncertainty of human expressions is well modeled in the fuzzy sets, and a pseudo-verbal representation, similar to an expert's formulation, can be achieved.

During fuzzy reasoning process, the input and output values of the neural network are generally fuzzified into linguistic terms so that fuzzy rules can be developed. The method of obtaining the corresponding output membership values from the "fired" fuzzy rule is called fuzzy logic reasoning. Many reasoning strategies have been developed, including Sup-bounded-product [18], Super-drastic-product [17,18,19], Sup-min [13], and Sup-product [11]. Since it is not the intention of this paper to present a review of fuzzy logic reasoning strategies, the mentioned reasoning strategies are not further explained in this paper. In this paper, the Sup-product strategy is adopted due to its simplicity and relatively less calculation time.

After the fuzzification process with the generation of fuzzy rules, it is necessary to have a defuzzification process. The defuzzification process is a process of mapping from a space of inferred fuzzy control results to a space of non-fuzzy control action in a crisp form. In fact, a defuzzification strategy is aimed at generating a non-fuzzy control action that best represents the possibility distribution of the inferred fuzzy control results. The Mean of Maximum (MOM) and Centre of Area (COA) are two common defuzzification methods in fuzzy control systems, and the latter method is selected in this neural-fuzzy model to defuzzify the reasoned fuzzy output (the parameters value). Proposed parameter change is carried out and the dimensional outcome, resulting from the change is checked against the expected dimension.

5 Cross Platform Intelligent Information Infrastructure

The real challenge for the implementation of information infrastructure in enterprises is to achieve *seamless* data interchange in the sense that data from various sources with dissimilar formats can integrate directly with the database system of any individual companies in a fully automatic way, i.e. without any human intervention. In this respect, data conversion and mapping of data-fields to match the formats of various enterprises are the pre-requisites of meeting this data integration criterion.

Since 1996, the new eXtensible Markup Language (XML) has been developed for overcoming the limitations of HTML, and has received worldwide acceptance in terms of data interchange in Internet-based information systems. In XML files, the content of data and the presentation of data are separated. Thus, various formats of data speci-

fied by XML are able to be transferred through the Internet and used on different platforms [16]. In brief, XML is a subset of Standard Generalized Markup Language (SGML) which is an international standard (ISO 8879) for defining and representing documents in an application-independent form [25].

In order to realize the potential of information infrastructure as an effective data processing gateway, data mining technology needs to be incorporated. In brief, data mining is seen as a technological approach that provides sophisticated analysis based on a set of complex data, enabling the management of different data formats in various database systems. The shared data access interface of data mining tools will enable exchange of data as well as results among various computer systems. The typical example of data mining tool is On-Line Analytical Processing (OLAP) that provides a service for accessing, viewing and analyzing on large volumes of data with high flexibility and performance. The essential characteristic of OLAP is that it performs a numerical and statistical analysis of data which are organized in the form of multi-dimensions (as opposed to the two-dimensional format of traditional relational data tables).

Cross platform intelligent information infrastructure allows users to access and retrieve data of any database format from distributed sources, followed by the automate assimilation of the data in the user's own database system. The information infrastructure also allows collaboration and data sharing and provides users with the ability to generate business transactions, reports and dynamic updates through the common interface. Based on the robust XML data interchange open standard, the information infrastructure can integrate virtually with all existing systems and database in real time and distrbute data automatically to the required destinations. In order to enable users to access timely information efficiently, an OLAP tool with intelligent functionality such as rule-based reasoning is required.

Fig. 5. shows the configuration of the information infrastructure which is a client-server system. The clients are users' (business partners, customers, employees) desktops who connect to the common interface using Internet browers. Thus, users can quickly access and retrieve all information and data needed through the information infrastructure, i.e. the Internet server with the XML and the intelligent OLAP facilities.

Currently in most client-server systems, the data requested from the client side is done by human users, such as requesting news on the Internet server or filling in the order forms. When the request originates from a database or a program on the client side, i.e. a database on the client side connects to the database on the server side, a standard data interchange format is needed. At this stage, a common data format such as the XML seems to be the right choice to meet the requirement. In order to create the XML data file, a XML translator is built in the information infrastructure so that different database systems can achieve data interchange among themselves.

Other than the data interchange through the information infrastructure, users may intend to perform data analysis such as statistical analysis. This task is normally referred as data mining. OLAP technology provides high flexibility and performance for the data mining and uses multidimensional data model for users to formulate complex queries and to arrange the query result on a report. The multidimensional data model organizes data into a hierarchy that represents levels of details on the data. Thus the

data structure of multidimensional data model is clearer than the relational database which is the structure of most current databases. In the information infrastructure, a multidimensional database is implemented for the OLAP application.

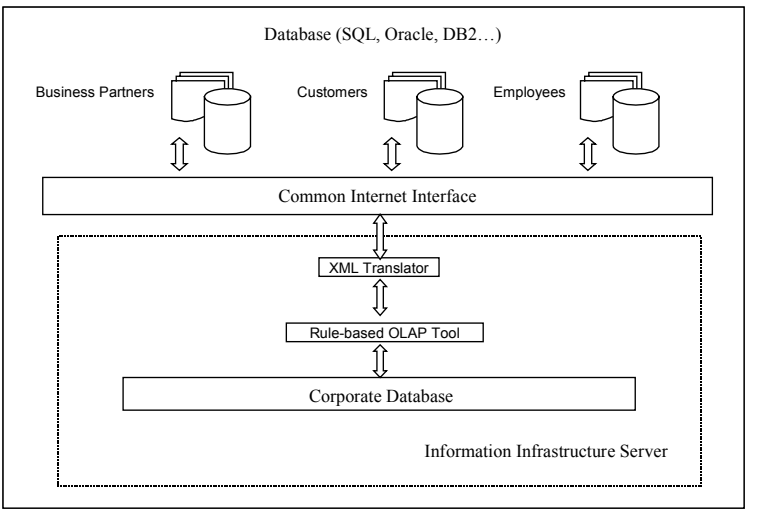

Fig. 5. Outline of Cross Platform Intelligent Informaton Infrastructure

Query on the large volume of data from the database by the OLAP tool is a complicated analysis and may not aggregate enough data. In the cross platform intelligent information infrastructure, the rule-based reasoning facility is incorporated to support the OLAP tool. The tasks of the rule-based reasoning are (i) to assist the setup of a corporate database, (ii) to help the extraction of useful information from the database, and (iii) to provide suggestions of efficient OLAP query methods. Thus, rule-based OLAP provides users with the access to the knowledge hidden in the databases, thereby discovering the trends and relationships in the data in order to make predictions using that information.

The rule-based reasoning facility includes knowledge base and inference engine. It retrieves knowledge (stored in the knowledge base) and uses the forward and backward inference logic to produce the actions (provide suggestions). In other words, it supports decision-makers (users) to get as much useful information as they can from their valuable knowledge base. In this respect, the rule-based reasoning can be deployed to enhance the intelligence level of the OLAP application.

The knowledge base of the rule-based reasoning facility consists of declarative knowledge and procedural knowledge [10]. The declarative knowledge contains what is currently known about the data interchange and process in the information infrastructure. The procedural knowledge is developed from repeated data interchange and process experience, and usually relates specific conditions with their likely outcomes, consequences, or indicated actions.

6 Conclusion

After a period of increasing professionalism in almost all areas of research and technology, the new era is to embrace the interaction of different approaches, to form multi-functional disciplines. The information infrastructures introduced above demonstrates the benefits of using combinations of technologies to form an integrated system, which capitalize on the merits and at the same time offset the pitfalls of the involved technologies. Further research on the infrastructure framework for knowledge discovery particularly the seamless integration of the technologies is needed in order to ensure and sustain the benefits of the infrastructures.

References

[1] Berson, A. and Smith, S.J. (1997), *Data Warehousing, Data Mining, & OLAP*, McGraw-Hill, New York.

[2] Buchanan, B. and Shortliffe, E.H. (1989). *Rule-based expert systems: the MYCIN experiments of the Stanford Heuristic Programming Project*. Addison-Wesley series in artificial intelligence. Reading, Mass.: Addison-Wesley.

[3] Burns R. (1997). Intelligent manufacturing. *Aircraft Engineering and Aerospace Technology*; 69(5): 440-446.

[4] Chiueh T. (1992) Optimization of Fuzzy Logic Inference Architecture, Computer, May; 67-71.

[5] Driankov D, Hellendoorn H, Reinfrank M. (1996) An Introduction to Fuzzy Control. Springer; 149-163.

[6] Erik, T., George, S. and Dick, C. (1999), *Microsoft OLAP Solutions*, John Wiley & Sons, New York.

[7] Haykin, S. (1994), *Neural networks, a comprehensive foundation*, Macmillan College Publishing Company.

[8] Haykin S. (1999). *Neural network, a comprehensive foundation*. 2nd edition. Upper Saddle River, N.J.: Prentice Hall.

[9] Herrmann, C.S. (1995), A hybrid fuzzy-neural expert system for diagnosis, *Proceedings of International Joint Conference on Artificial Intelligence*, pp. 494-500.

[10] Inference Corporation, 1992, *ART-IM 2.5 Reference Manuals* (Los Angeles).

[11] Kaufman A. (1975). *Introduction to theory of fuzzy subsets*. New York, Academic.

[12] Leung KS, and Lam W. (1988). Fuzzy Concepts in Expert Systems. *IEEE,* September, 43-56.

[13] Mamdani EH. (1974) Applications of fuzzy algorithms for control of a simple dynamic plant. *Proceedings of IEEE, 1974*; 121: 1585-1588.

[14] Merwe, J.v.d. and Solms, S.H.v. (1998), Electronic commerce with secure intelligent trade agents, *Computers & Security*, Vol. 17, pp. 435-446.

[15] Michael, L.G. and Bel, G.R. (1999), Data mining - a powerful information creating tool, *OCLC Systems & Services*, Vol.15, No. 2, pp81-90.

[16] Microsoft Corporation, 2000, *Microsoft BizTalk jumpstart kit*, Feb.
[17] Mizumoto M, Fukami S, Tanaka K. (1979). Some Methods of Fuzzy Reasoning. Advances in Fuzzy Set Theory and Applications. North-Holland, Amsterdam; 117-136.
[18] Mizumoto M. (1981) Note on the arithmetic rule by Zedeh for fuzzy reasoning methods. *Cyben System*; 12: 247-306.
[19] Mizumoto M. (1990) Fuzzy controls by product-sum-gravity method. In Advancement of fuzzy theory and systems in China and Japan, *Proceeding of Sino-Japan Joint Meeting on Fuzzy Sets and Systems Oct. 15-18*, Beijing, China, International Academic; 1-4.
[20] New Era of Networks, Inc., (2000), Powering the new economy.
[21] Nguyen HT. (2000) *A first course in fuzzy logic*. 2nd edition. Boca Raton, Fla: Chapman & Hall/CRC.
[22] Orchard A. (1994) *FuzzyCLIPS Version 6.02A User's Guide*. National Research Council. Canada.
[23] Peterson, T. (2000) *Microsoft OLAP unleashed*, 2nd edition, Sams Pubishing, Indianapolis.
[24] Robert S.C., Joseph A.V. and David B. (1999), *Microsoft Data Warehousing*, John Wiley & Sons.
[25] Salminen, A., Lyytikäinen, V. and Tiitinen, P., (2000), Putting documents into their work context in document analysis, *Information Processing & Management* 36, Issue 4, July 1, 623-641.
[26] Tandem Computers Incorporated (1997), Object Relational Data Mining Technology for a Competitive Advantage (White Paper), *Decision Support Solutions*.
[27] Thomsen, E. (1999), *Microsoft OLAP solutions*, J. Wiley, New York.
[28] Whalen T, Schott B. (1983) Issues in Fuzzy Production Systems. *International Journal of Man-Machine Studies*; 19:57.
[29] Zadeh, F. (1996) Fuzzy sets, fuzzy logic, and fuzzy systems: selected papers. Singapore, World Scientific.
[30] Zadeh, F. (1965) Fuzzy sets. *Information and Control* 8:338-53.
[31] Zadeh, F. (1993) The role of fuzzy logic and soft computing in the conception and design of intelligence systems. Klement, Slany.
[32] Zahedi F. (1991) An introduction to neural network and a comparison with artificial intelligence and expert systems. *Interfaces*; 21(2): 25-28.
[33] Zahedi F. (1993) Intelligent Systems for Business: Expert Systems with Neural Network, Wadsworth, Belmont, CA.