6

Impacts of data mining technology on product design and process planning

Cihan H Dagli and Hsi-Chieh Lee
Smart Engineering Systems Laboratory
Department of Engineering Management, University of Missouri-Rolla
Rolla Missouri 65409-0370
dagli@umr.edu
Phone: 573 341-4374
Fax: 573 341-6567

Abstract

Recent advances in computers and networking technologies and a fast-growing internet community created immense distributed data bases located miles away and having a capability to be updated continuously without the knowledge of the possible and prospective users. The ability to collect and store all kinds of data have outpaced the capabilities of individuals to analyze, summarize, and extract "knowledge" from them. Traditional methods of data analysis, based mainly on the analysts dealing directly with the data, is no longer the best alternative to be used. Although the database technology provided the basic tools for efficient storage and lookup for large data sets, the issues of how to enable engineers to understand large bodies of data remains a difficult problem. Recently, data mining approaches based on artificial neural networks, fuzzy logic, machine learning, statistics, expert systems, and data visualization are creating new intelligent tools for automated data mining and knowledge discovery.

All these changes will have a profound impact on current practices used in manufacturing. The way bills of materials are created, products designed and process plans generated will be definitely different with the availability of this new technology. In this paper the nature of these changes and their implication on current practices will be discussed in reference to an intelligent data mining system being developed in the smart engineering system design laboratory.

Computer Applications in Production and Engineering. F. Plonka and G. Olling (Eds.)

1 INTRODUCTION

The advances of computer and networking technologies and fast-growing Internet community have brought a data glut problem to the world of science, business and government. The capabilities for collecting and storing data of all kinds have far outpaced the abilities to analyze, summarize, and extract "knowledge" from this data. The Manufacturing sector is not immune to this change.

Traditional methods of data analysis, based mainly on the individual dealing directly with the data are becoming obsolete. While database technology has provided us with the basic tools for the efficient storage and lookup of large data sets, the issue of how to help humans understand and analyze large bodies of data remains a difficult and unsolved problem. To deal with the problem, a new generation of intelligent tools for automated data mining and knowledge discovery is needed. This need has been recognized by researchers in different areas, including machine learning, statistics, intelligent databases, neural networks, fuzzy systems, expert systems, and data visualization. In this paper the impact of these new technologies on manufacturing practices, namely, product design and process planning is discussed.

The Internet is basically a network of networks. It currently connects millions of networks to allow users to globally share information and computer resources. With the existence of the Internet, an user can share virtually anything that can be stored in a file. Internet communication is possible among networks on different platforms and in different environments. This capability of exchanging data dynamically is in part due to the development of communication protocols. Protocols are agreed-upon standards for exchanging data, and enable computing devices to communicate among various networks.

The Transmission Control Protocol and the Internet Protocol (TCP/IP) were developed in the 1960s to provide a communication link, even if some of the connecting links between the devices were to fail. In 1969, the Department of Defense began using ARPANET, the first network based on the protocol technology. ARPANET initially connected four supercomputers. In the 1970s, educational and research institutions began to connect to ARPANET to create a community of networks. In the late 1970s, TCP/IP became the official protocol to use on the Internet. During the 1980s, the U.S. National Science Foundation replaced ARPANET with a high-speed network. This is the network that now serves as the backbone for the Internet today. When ARPANET was first used in 1969, it consisted of only 213 registered hosts. By 1986, there were over 2300 host computers. In the early 1990s, the U.S. National Science Foundation transferred the maintenance and funding of the Internet to private foundations and corporations. Today, the Internet has several million host computers worldwide. The development of other protocols and other technologies, such as the World Wide Web, has contributed to this growth.

The World Wide Web (WWW) is a distributed hypermedia system for information discovery, retrieval, and collaboration. It was created by scientists at CERN who wanted to share and gain access to research information through a

common interface. By using a common interface, researchers no longer had to perform the many steps necessary to gain access to the different available Internet services. More and more people who use the Internet have seen the value of using a common interface (i.e. web browser). In just three years since it was introduced, the Web has grown to include users from all ages and vocations. It has proven its usefulness for browsing large, distributed document structures. However, as the amount of information available through the World Wide Web increases, it becomes more and more important to provide additional tools and techniques for finding servers of documents which contain relevant information on one's special interest. The main difference between a hypertext network and conventional linear text is that in a hypertext system, navigation is up to the user. The HyperText Transfer Protocol [1] is used by web servers to communicate with each other and a variety of client applications, such as, FTP, Gopher, and WAIS.[2,3].

Manufacturing systems of the twenty-first century need to be able to use this dynamic distributed data base environment, and change their product design and process planning practices in time. Hence, there is a need to design "smart" systems for this purpose that can interact with their environment, namely, continuously changing distributed manufacturing information base residing on the interconnected computers of the world and adapt to the changes in time and space by their ability to manipulate the environment through self-awareness and perceived models of the world based on both quantitative and qualitative information. This need will integrated base functions of manufacturing, product design, process planning and control even further to be able to respond to global customer's changing requirements. The emerging technologies of artificial neural networks, fuzzy logic, evolutionary programming and data mining will provide essential tools for designing these smart manufacturing systems. In the following section a brief description of these technologies is given.

2 EMERGING TECHNOLOGIES

2.1 Neural Networks

Neural Network models have been studied and used extensively in the last decade in order to achieve human-like intelligence. The earlier works by McCulloch and Pitts [4], Hebb [5], Rosenblatt [6], and Widrow [7] as well as the more recent works by Feldman [8], Grossberg [9], Hopfield [10, 11, 12], Rumelhart and McClelland [13], Sjnowski [14], and others [15- 20] have brought the world many useful applications. Typical applications include: classification, decision making, financial analysis, medical diagnostics, optimization, pattern recognition, process control, robotics and automation, signal processing, and targeted marketing, time series prediction.

Artificial Neural Networks are the mathematical models, which represent the biological process of the human brain. In an Artificial Neural Network, there are three main components: neurons, interconnections, and learning rules.

Neurons

The neuron is a simple device which approximates the function of the fundamental unit of the biological nervous systems. It receives and processes signals from either other neurons or the outside environment, then continually passes its output to the next level neurons. Each neuron can receive many input signals simultaneously, but there is only one output signal which depends on the input signals, weights of connections, threshold, and the activation function.

Neurons can be classified into three types based on their inputs and outputs: input, output and hidden neurons. Input neurons are the ones that actually receive input from the environment. Output neurons are those that send the signals out of the system, and neurons, which have inputs and outputs within the system, are called hidden neurons.

The purpose for including the activation function in the neuron is to confine the neuron's output to a pre-specified range.

Interconnections

The interconnection is a part of the network architecture, which propagates signals in a single direction from one neuron to the others, or even to itself. There is a weight value assigned to each connection.

There are three different kinds of connections that link neurons in a network.

- Intrafield connections: Connect neuron in the same layer.
- Interfield connections: Connect neuron in different layers.
- Recurrent connections: Connect neuron to its self.

Learning Rules

Learning in an artificial network is considered to as a change in the weight matrix. It can be categorized into two groups: supervised and unsupervised learning.

Supervised learning uses the data set that contains input vectors and corresponding output vectors to train the network. The weight matrix of the network is updated as long as the total network error is greater than ε (acceptable range of error). Such updating techniques as error-correction learning, reinforcement learning and stochastic learning are always included in this type of learning. *Error-correction learning* adjusts the whole weight matrix in proportion to the difference between the desired and the actual values of the output neuron. *Reinforcement learning* is a technique in which weights are reinforced for properly performed actions and punished for inappropriate ones. *Stochastic learning* makes a random change in the weight matrix then determines resultant energy of the network. If the resultant energy is lower than the previous one, the change is accepted, otherwise the change is considered under a pre-chosen probability distribution.

Unsupervised learning does not incorporate external teacher. It relies only on local information and internal control.

2.2 Fuzzy Logic

Multivalued logic was first introduced in the 1920s and 1930s as a result of logical paradoxes. In 1965, Lofti Zadeh [21] published the paper "Fuzzy Sets" which formally developed multivalued set theory. This was thought of as the first time that the term "fuzzy" was introduced into the technical literature, and which inaugurated a second wave of interest in multivalued logic system. The recent new theory and application in fuzzy logic led to the current third wave of interest in fuzzy systems. Fuzziness measures the degree that an event may occur or the degree that an entity may be classified as something. Fuzzy logic systems permit variables to belong to more than one set of class. [22]. The fuzzy set concept used in intelligent data mining system described in this paper was introduced by Zadeh [21, 23] and summarized by Kosko [24].

2.3 Data Mining

Data mining (also known as Knowledge Discovery in Databases- KDD) has been defined as "The nontrivial extraction of implicit, previously unknown, and potentially useful information from data" [25]. It uses machine learning, statistical and visualization techniques to discovery and present knowledge in a form which is easily comprehensible to humans. There might be valuable information in your data, but you simply cannot see it. It could not be as profound as a new law of nature. But no human who has looked at your data has seen this hidden information which might be important. How can people find it? Data mining lets the power of computers do the work of sifting through the vast data stores.

In recent years, data mining has attracted the interest of many researchers due to its importance of finding knowledge in growing size of databases [26-36]. Data mining deploys a variety of algorithms. Generally, the more algorithms in use, the higher the likelihood of accurate results. There are no certain rules and tools. As a result, current data mining researchers are fielding architectures that combine these techniques: neural networks, induction, association, fuzzy logic, statistical, and visualization.

Gerber [37] summarized the commonly used approaches into the following four categories:

- Predictive modeling: In OLAP (On-Line Analytical Processing), it uses deductive reasoning; in data mining, it uses inductive reasoning. Predictive modeling can be implemented in a variety of ways, including neural networks or induction algorithms.
- Database segmentation: The automatic partitioning of a database into clusters. It generally uses statistical clustering in its implementation.
- Link analysis: Identifying connections between records, based on association discovery and sequence discovery.

- Deviation detection: The detection of an explanation for why records cannot be put into segments. This can be implemented via different kinds of statistics.

Generally, data-mining process includes the following sub-tasks:

- Preprocess data: It includes the data collection, data cleaning, and data storing.
- Search for patterns: This is usually the most crucial part of the data-mining process. A number of tools are being used for this purpose such as queries, rules, neural networks, machine learning, and statistics.
- Analyst reviews output: The output of the previous sub-task is investigated here to decide whether to report the findings or to perform a revised pattern search.
- Report findings: The findings will be further interpreted and response accordingly.

All of these new technologies are incorporated into the intelligent data mining system that is under development.

3 IMPACTS ON MANUFACTURING PRACTICES

Availability of adaptive data mining system can have profound impact on product design and process planning. These can be summarized briefly as described in the following paragraphs

3.1 Impact on Product Design

Engineering design is essentially the process of converting the desires and needs of the customer into detailed specifications for a useable end product. It is a highly knowledge intensive and time-consuming activity which requires a great many decisions and judgments on the part of the human designer. It is also an imprecise art, since each designer differs in background, experience, preferences, and formal training. The more knowledge and experience the designers have, the better the chances for generating creative designs. Final product design impacts sixty to eighty-five percent of the total manufacturing cost. Hence, through analysis of design alternatives is essential prior to the finalization of the product design process. This is not an easy task as the number of possible design solutions increases exponentially as new functional requirements and manufacturing constraints are introduced. The NP-complete (non polynomial in time characteristic of the design problem necessitates different search strategies in selecting the final design creating a wide variety of approaches adopted by researchers. How does an artifact get designed? How do fuzzy mental images and abstract concepts get converted to the crisp design: The process is not well understood at this time. Each researcher approaches the problem in a different way in an effort to generate the best design satisfy functional requirements.

Suh defines design as the culmination of synthesized solutions in the form of products, software, processes, or systems by the appropriate selection of design parameters that satisfy perceived needs through the mapping from functional requirements in the function domain to design parameters in the structure domain. This mapping process is not unique, and more than one design may result from the generation of design parameters that satisfy the functional requirements. Therefore, there can be an infinite number of feasible design solutions and mapping techniques. Suh's design axioms provide the principles the mapping technique must satisfy in regard to input requirements in order to produce a good design and provide a framework for comparing and selecting designs. The axioms are independence and information. The independence axiom states that in an acceptable design, the design parameters and the functional requirements are related in such a way that a specified design parameters can be adjusted to satisfy its corresponding functional requirements without affecting other functional requirements. The information axiom states that the best design is a functionally uncoupled design that has a minimum information content.

Many researchers have been trying to develop a scientific theory or at least a mode of design. Their perspectives and views on the nature of design can be visualized as a feed-back loop of synthesis, evaluation, and analysis. The general model of the design process is inherently iterative and requires improvements until requirements are satisfied.

Product design goes through all the above stages. Intelligent data mining system can provide valuable information to the designer in all stages as it has the capability to predict the need of the designer and receive relevant information from the interconnected computers of the global village. This capability will impact four basic distinct aspects of product design:

- Problem definition from fuzzy sets of facts and myths into a coherent statement of the question
- Creative process of devising a proposed physical embodiment of solutions
- Analytical process of devising a proposed physical embodiment of solutions
- Ultimate check of the fidelity of the design product to the original perceived needs

3.2 Impact on Product Design

Manufacturing planning, process planning, material processing, process engineering and machine routing are some of the names given to the topic referred to here as process planning.

Process planning is the act of preparing detailed operation instructions to transform an engineering design to a final part. It is the critical bridge between design and manufacture. Design information can only be translated through process planning into manufacturing language. The detailed plan contains the route, processes, process parameters, machines, and tools required for production. The process plan provides the instructions for production of the part. These instructions dictate the cost, quality and rate of production. Therefore process

planning is of utmost importance to the production system. In general, a process plan is prepared using the available design data, and manufacturing knowledge.

Process planning is a data intensive activity that requires extensive search. Intelligent data mining system due to its ability to predict the need of the process planner and locate the source of information and knowledge to restrict the search can provide an important input in generating process plans. It is very difficult to envision the impact of this capability in manufacturing practices.

4 INTELLIGENT DATA MINING SYSTEM

4.2 The Model

The term "Cyberspace" has existed for decades, while the term "World Wide Web" had not been known until recently. Cyberspace is a unified conceptualization of space spanning the entire Internet. It is a spatial equivalent of the World Wide Web. There is only one Cyberspace, just on an irregular network topology. It has brought the information from a distributed environment into a global information universe. As the number of web servers increases, surfing the Web is increasingly difficult. The proposed model, Cyber Agent, is seeking a computational model which can efficiently implement high-level intelligent processes instead of seeking to model the detailed biological aspect of the human brain.

The Cyber Agent (Fig. 1) is a smart engineering system built to help users search and organize the information. It contains two major subsystems, namely, WebTracer and WebOrganizer. They adapt behavior dynamically according to the environment and the special requirements of each individual. WebTracer is the wavefront of the CyberAgent while WebOrganizer is the brain of the CyberAgent. They are implemented using PERL and Common Gateway Interface (CGI) [38].

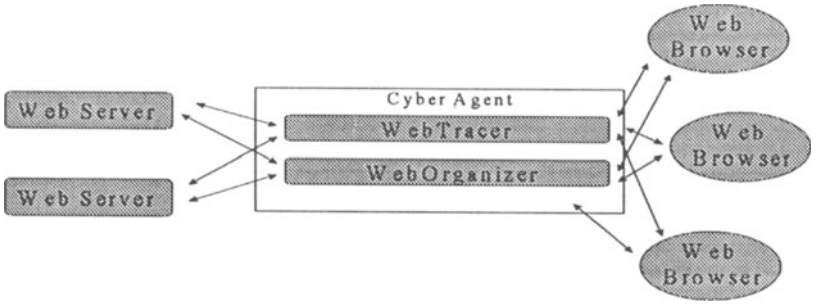

Figure 1: The CyberAgent Model

4.2 WebTracer

WebTracer (see Fig. 2) is the wavefront of the Cyber Agent which obtains the information from Cyberspace. While there are a number of search engines available which hold a tremendous amount of data, WebTracer is not meant to be another database-base search engine. Instead, it resides on top of the existing search engines, web robots, web spiders, and web wanderers [39]. This speeds up the searching process and eliminates the requirement for huge storage locally. It incorporates WAIS's [3] ability which answers Z39.50 information requests to search several databases at once and lets the users select those databases. A hybrid system consisting of neural networks and fuzzy associative memory is used for conceptual search and approximate string matching. It improves queries based on the user's feedback. It creates links dynamically to the user's particular interests using HyperText Markup Language (HTML) [40].

The beauty of the World Wide Web is the wide availability of resources. To utilize the resources in Cyberspace, WebTracer first consults the online libraries which contain dictionaries, thesauruses, and encyclopedias for the keywords that the user provided and then launches the existing search engine to search the related information obtained from the online libraries.

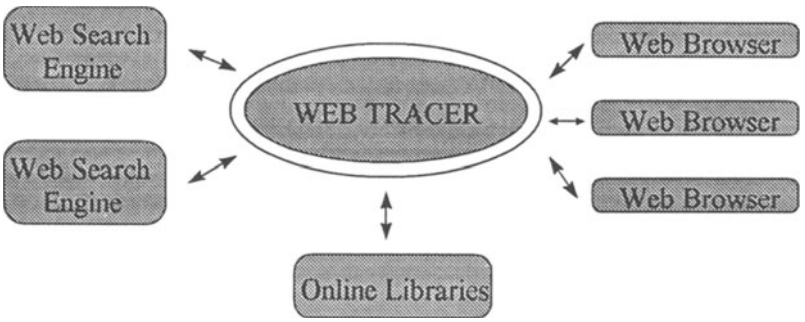

Figure 2: The WebTracer Diagram

4.3 WebOrganizer

The popularity of web servers and web browsers has brought the information from a distributed environment into a global information universe which is available at one's fingertips. As more information is available in Cyberspace, the users will find more information interesting to them. As the user surfs the web and adds bookmarks, it is increasingly difficult for them to find their desired information from their bookmarks.

The WebOrganizer (see Fig. 3) is the brain of the CyberAgent. It serves as a housekeeper of the web pages and bookmarks. Since the information in Cyberspace and the distributed network environment is maintained by many people, documents may be moved or deleted, referenced information may change, and the hyper-links (Uniform Resource Locators, URL[41]) may be broken. Like

MOMspider [42], WebOrganizer maintains the hyper-links by transversing in Cyberspace. It checks the existence of hyper-links, moved documents, and recent modification dates. Besides, WebOrganizer can also expand the bookmarks by referring to the existing hyper-links in it. It also has the capability to organize the bookmarks in a variety of ways specified by the user. WebOrganizer consults WebTracer for the validity of hyper-links, and recent modification dates of hyper-links.

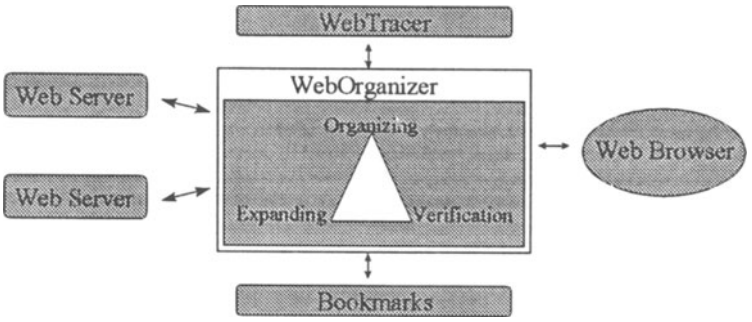

Figure 3: The WebOrganizer Diagram

5 REMARKS

Manufacturing engineers of the future need to understand and use large bodies of manufacturing data distributed globally and should be able to make sound product design and process planning decisions. This is possible through smart data base systems that can interact with their environment and adapt to changes both in space and time. Emerging technologies of artificial neural networks, fuzzy logic and evolutionary programming are providing the necessary foundation to design such systems as described by the intelligent data mining system.

6 REFERENCES

[1]"Hypertext transfer protocol,"
URL:http://www.info.cern.ch/hypertext/WWW/Protocols/HTTP/HTTP2.html.

[2] Kahle, B. (1991) *An information System for Corporate Users*, WAIS. Thinking Machines Corporation.

[3] "Wide-area information system," *URL:http://www.wais.com/z3950.html*

[4] McCulloh, W. S. and Pitts W. (1943) "A logical calculus of the ideas imminent in nervous activity", *Bulletin of Mathematical Biophysics*, **5**, 115-132.

[5] Hebb, D. O. (1949) *The Organization of Behavior.* Jan Wiley & Sons, New York.

[6] Rosenblatt, R. (1959) *Principles of Neurodynamics.* Spartan Books, New York.

[7] Widrow, B. and Hoff M. E. (1996) "Adaptive Switching Circuits," *1960 IRE WECON* **Conv. Record Prt. 4**, 96-104.

[8] Feldman, J. A. and Ballard D. H. (1982) "Connectionist models and their properties," *Cognitive Science*, **6**, 205-254.

[9] Grossberg, S. (1996) The Adaptive Brain I: *Cognition, Learning, Reinforcement, and Rhythm*, and The Adaptive II: *Vision, Speech, Language, and Motor Control.* Elsevier/Nth-Holland, Amsterdam.

[10] Hopfield, J. J. (1982) "Neural networks and physic systems with emergent collective computational abilities," *Proc. Natl. Acad. Sci. USA*, **79**, 2554-2558.

[11] Hopfield, J. J. (1984) "Neurons with graded response have collective computational properties like those of two-state neurons," *Proc. Natl. Acad. Sci. USA*, **81**, 3088-3092.

[12] Hopfield, J. J. and Tank, D. W. (1986) "Computing with neural circuits: A model," *Science*, **233**, 625-633.

[13] Rumelhart, D. E., Hinton, G. E. and Williams, R. J. (1986) "Learning internal representations by error propagation," *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, **1**

[14] Sejnowski, T. and Rosenberg, C. R. (1986) "Nettalk: A parallel network that learns to read aloud," *Johns Hopkins Univ. Technical Report*, **JHU/EECS-86/01**.

[15] Lippmann, R. P. (1987) "An introduction to computing with neural nets," *IEEE ASSP Mag.*, **4(2)**, 4-22.

[16] Kohohen, T. (1989) *Self-Organization and Associative Memory.* Springer-Verlag, 3rd ed.

[17] Carpenter, G. A. and Grossberg, S. (1987) "A massively parallel architecture for a self-organization neural pattern recognition machine," *Computer Vision, Graphics and Image Processing*, **37**, 54-115.

[18] Carpenter, G. A., Grossberg, S. and Rosen, D. B. (1991) "Fuzzy art: Fast stable learning and categorization of analog patterns by an adaptive resonance system," *Neural Networks*, **4**, 759-771.

[19] Smith, M. (1993) *Neural Networks for Statistical Modeling*. Van Nostrand Reinhold.

[20] Lee, H. C., Dagli, C. H. and Ercal, F. (1994) "Neuro-fuzzy approach for data clustering: A prototype architecture," *Submitted to Artificial Neural Networks In Engineering*.

[21] Zadeh, L. A. (1965) "Fuzzy sets," *Information and Control*, **8**, 338-353.

[22] Lawrence, J. (1992) *Introduction to Neural Networks and Expert Systems*. California Scientific Software.

[23] Zadeh, L. A. (1983) "A computational approach to fuzzy quantifiers in natural languages," *Computers and mathematics*, **9(1)**, 149-184.

[24] Kosko, B. (1992) *Neural Networks and Fuzzy Systems; A dynamical Systems Approach To Machine Intelligence*. Prentice-Hall.

[25] Frawley, W., Piatetsky-Shapiro, G. and Matheus, C. (1991) "Knowledge discovery in databases: An overview," *AI Mag.*, 213-228, Fall(Autumn) 1992. Reprint of the introductory chapter of Knowledge Discovery in Databases collection, AAI/MIT Press.

[26] Agrawal, R., Imielinski, T. and Swami, A. (1993) "Database mining: A performance perspective," *IEEE Transactions on Knowledge and Data Engineering*, **5**, 914-925.

[27] Kaufman, K. A., Michalski, R. S. and Kerschberg, L. (1991) *"Mining for knowledge in databases: Goals and general description of the inlen system,"* in Knowledge Discovery in Databases (G. Piatetsky-Shapiro and W. J. Frawley, eds.), Menlo Park, California: AAAI Press / The MIT Press, 1st ed.

[28] Michalski, R., Kerschberg, L. and Kaufman, K. (1992) "Mining for knowledge in databases: The inlen architecture, initial implementation and first results," *Journal of Intelligent Information Systems*, 85-113.

[29] "Data mining: Intelligent technology gets down to business," *PC AI*, Nov-Dec 1993

[30] Agrawal, R., Imielinski, T. and Swami, A. (1993) "Database mining: A performance perspective," *IEEE transactions on knowledge and data engineering*, **5(6),** 914-925.

[31] Michalski, R. S., Kerschberg, L. and Kaufman, K. A. (1992) "Mining for knowledge in databases: the INLEN architecture, initial implementation and first results," *Journal of Intelligent Information Systems*, **1**, 85-113.

[32] Al-naemi, S. (1992/3) "*Temporal aspects in data mining*," tech. Rep., Computer Science Department, Univ. of Birmingham.

[33] "Getting to grips with arrears: 'data mining' systems at the Leeds," *Expert Systems*, **11**, 122-124, 1994.

[34] Holsheimer, M. and Siebes, A. P. (1994) "*Data mining: the search for knowledge in databases*," Tech Rep. CS-R9406, CWI.

[35] Zytkow, J. M. and Baker, J. (1991) "*Interactive mining for regularities in databases*," 31-35, Menlo Park, California: AAAI Press.

[36] Holsheimer, M. and Siebes, A. P. (1994) "*Data mining: the search for knowledge in databases*," Tech. Rep. CS-R9406, CWI.

[37] Gerber, C. (1996) "Excavate your data," *Datamation*, **42(9).**

[38] McCool, R. (1993-1994) "*The Common Gateway Interface*," National Center for Supercomputer Applications, Univ. of Illinois at Urbana-Champaign.

[39] Koster, M. "World Wide Web Robots, Wanderers, and Spiders" *URL:http://www.infor.webcrawler.com/mak/projects/robots/robots.html*

[40] Berners-Lee, T. (ed.), "Hypertext Mark-up Language," CERN *URL:http://www.infor.cern.ch/hypertext/WWW/MarkUp/MarkUp.html*

[41] Berners-Lee, T. (1993) "Uniform Resource Locators," *Internet Engineering Task Force Draft*, CERN, Geneva, Switzerland.

[42] Fielding, R. T. (1994) "Maintaining Distributed Hypertext Infostructures: Welcome to MOMspider's Web," University of California at Irvine, *URL:http:///www.ncsa.uiuc.edu/demonweb/html-primer.html*