International Journal of Computer Engineering and Applications,
Volume XII, Special Issue, May 18, www.ijcea.com ISSN 2321-3469

IMPLEMENTATION OF DATA MINING TECHNIQUES FOR EXTRACTION OF KNOWLEDGE MANAGEMENT

Yethiraj N G[1] Sumanth S[2]
[1] Assistant Professor, Department of Computer Science
Maharani's Science College for Women, Palace Road, Bangalore -560001
[2]Assistant Professor, Department of Computer Science
Smt. VHD Central Institute of Home Science, Seshadri Road, Bangalore-560001

ABSTRACT:

Knowledge Management systems benefit corporations that take advantage of the Artificial Intelligence technology. As enterprises are being driven toward KM systems to meet competitive pressures and create value, they are increasingly finding that these systems can facilitate reuse of existing knowledge and create new knowledge in an effort to allow better decision making process. In this paper we argue that Data mining and Data warehousing with allied AI concepts can make a significant contribution to knowledge management initiatives.

Keywords: Data Warehouse, Data Mining (DM), Artificial intelligence (AI), Ontology, Knowledge Management (KM), and Information Technology (IT)

[1] INTRODUCTION

In the recent years the concept of knowledge management has become very popular especially in the software engineering environments. The value of Knowledge Management relates directly to the effectiveness [bel97a] [11][12] with which the managed knowledge enables the members of the

organization to deal with today's situations and effectively envision and create their future. Without on-demand access to managed knowledge, every situation is addressed based on what the individual or group brings to the situation with them. With on-demand access to managed knowledge, every situation is addressed with the sum total of everything anyone in the organization has ever learned about a situation of a similar nature. [bel97b] [12]

It has been argued that with organizations increasingly exposed to global competition in their business, knowledge has become critical for organizations to survive (Davenport and Prusak, 1998). The current emphasis in theory and practice of KM is to understand knowledge creation, transmission, storage and retrieval. Information technology (IT) is considered an important part of KM initiatives (KPMG, 2002). The general assumption behind it is that IT can positively affect knowledge sharing across the organization [2].

Many number of management theorists have contributed to the evolution of knowledge management, among them such notables as Peter Drucker, Paul Strassmann, and Peter Senge in the United States. Drucker and Strassmann have stressed the growing importance of information and explicit knowledge as organizational resources, and Senge has focused on the "learning organization," a cultural dimension of managing knowledge. Chris Argyris, Christoper Bartlett, and Dorothy Leonard-Barton of Harvard Business School have examined various facets of managing knowledge. In fact, Leonard-Bartonï¿½s well-known case study of Chaparral Steel, a company which has had an effective knowledge management strategy in place since the mid-1970s, inspired the research documented in her Wellsprings of Knowledge ï¿½ Building and Sustaining
Sources of Innovation (Harvard Business School Press, 1995). Everett Rogersï¿½ work at Stanford in the diffusion of innovation and Thomas Allen research at MIT in information and technology transfer, both of which date from the late 1970s, have also contributed to our understanding of how knowledge is produced, used, and diffused within organizations. By the mid-1980s, the importance of knowledge and its expression in professional competence as a competitive asset was apparent, even though classical economic theory ignores the value of knowledge as an asset and most organizations still lack strategies and methods for managing it. Recognition of the growing importance of organizational knowledge was accompanied by concern over how to deal with exponential increases in the amount of available knowledge and increasingly complex products and processes. The computer technology that contributed so heavily to superabundance of information started to become part of the solution, in a variety of domains. Doug Engelbartï¿½s Augment (for "augmenting human intelligence"), which was introduced in 1978, was an early hypertext/groupware application capable of interfacing with other applications and systems. Rob Acksynï¿½s and Don McCrackenï¿½s Knowledge Management System (KMS), an open distributed hypermedia tool, is another notable example and one that predates the WWW by a decade

Like water, this rising tide of data can be viewed as an abundant, vital and necessary resource. With enough preparation, we should be able to tap into that reservoir and ride the wave by utilizing new ways to channel raw data into meaningful information. That information, in turn, can then become the knowledge that leads to wisdom. Les Alberthal [alb95] [11]

In his seminal paper on the knowledge level Newell (1982) situates knowledge in the epistemological processes of an observer attempting to model the behavior of another agent:

"The observer treats the agent as a system at the knowledge level, i.e. ascribes knowledge and goals to it." (p.106) emphasizing that:

"The knowledge level permits predicting and understanding behavior without having an operational model of the processing that is actually being done by the agent." (p.108)

He defines knowledge as:

"Whatever can be prescribed to an agent such that its behavior can be computed according to the principle of rationality." (p.105) noting that:

"Knowledge is that which makes the principle of rationality work as a law of behavior." (p.125) and defining rationality in terms of the principle that:

"If an agent has knowledge that one of its actions will lead to one of its goals, then the agent will select that action." (p.102)

In the light of these analyses, Newell's arguments may be seen as stating that knowledge is a state variable imputed by a modeler in order to account for its behavior, and that the appropriate presuppositions for
modeling an agent are those of rational teleology, that it has goals and acts to achieve them. [4]

According to Mike Davidson [dav96], what's really important is:

- Mission: What are we trying to accomplish?
- Competition: How do we gain a competitive edge?
- Performance: How do we deliver the results?
- Change: How do we cope with change?

As such, knowledge management, and everything else for that matter, is important only to the extent that it enhances an organization's ability and capacity to deal with, and develop in, these four dimensions

Enterprise knowledge management entails formally managing knowledge resources in order to facilitate access and reuse of knowledge, typically by using advanced IT. KM is formal in that knowledge is classified and categorized according to a pre specified but evolving ontology into semi structured and structured knowledge and data bases. The overriding purpose of enterprise KM is to make knowledge accessible and reusable to the enterprise. Knowledge resources vary for particular applications and organizations, but they generally include manuals, letters, summaries of responses to clients, knowledge derived from work processes, customer information, intelligence of competitor, and news. There are wide range of technologies are being used to implement KM systems like E-mail; databases and data warehouses; group support systems; browsers and search engines; intranets and internets; expert systems and knowledge-based systems and intelligent agents. In AI, knowledge bases are generated for consumption by so-called expert and knowledge-based systems, where computers use rule inference to answer user questions. Although knowledge acquisition for computer inference is still important, most recent KM developments make knowledge available for direct human consumption or develop software that processes that knowledge. Historically, KM has been aimed at single group managers through what has been generally referred to as executive information system. An EIS contains a portfolio of tools such as drill-down access to data bases, news source alerts, and other information all aimed at supporting managerial decision making. More recently, however, KM systems are increasingly designed for entire organizations. If executives need access to information and knowledge, their employees are also likely to have an interest in and need for that information. In addition, KM technology is ideally suited for non management groups—such as customer support, where customer service requests and their solutions can be codified and entered into a database available to all customer service representatives. [3].

The principal purpose of data ware housing is to provide information to business users for strategic decision making. These users interact with the data warehouse using front-end tools, or by getting the required information through the information delivery system. Different types of users

engage in different types of decision support activities, and therefore require different types of tools including 4GL(Fourth Generation Language) , EIS (Executive Information System), Spreadsheets, OLAP (Online Analytical Processing) and data mining. [6][10].

The broader definition for Data mining given by Han and Kamber (2001): data mining is the process of discovering interesting knowledge from large amounts of data stored either in databases, data warehouses, or other information repositories. Knowledge we mean patterns, rules or relationships between data not easily identifiable using the cognition capabilities of a human being; that is, non-trivial, implicit and potentially useful information. [1]

2. DATA MINING – AN ITERATIVE PROCESS

DM is an iterative process within which progress is defined by discovery, through either automatic or manual methods. DM is most useful in an exploratory analysis scenario in which there are no predetermined notions about what will constitute an "interesting" outcome. DM is the search for new, valuable, and nontrivial information in large volumes of data. It is a cooperative effort of humans and computers. Best results are achieved by balancing the knowledge of human experts in describing problems and goals with the search capabilities of computers.

Prediction and description are the two primary goals of DM in practice. Prediction involves using some variables or fields in the data set to predict unknown or future values of other variables of interest. Description, on the other hand, focuses on finding patterns describing the data that can be interpreted by humans. Therefore, it is possible to put DM activities into one of the following two categories:

1) Predictive data mining, which produces the model of the system described by the given data set, or
2) Descriptive data mining, which produces new, nontrivial information based on the available data set. [1][8].

DM uses well-established statistical and machine learning techniques to build models that predict customer behavior. Statistics has its roots in mathematics, and therefore, there has been an emphasis on mathematical rigor, a desire to establish that something is sensible on theoretical grounds before testing it in practice. In contrast, the machine-learning community has its origins very much in computer practice. This has led to a practical orientation, a willingness to test something out to see how well it performs, without waiting for a formal proof of effectiveness. If the place given to mathematics and formalizations is one of the major differences between statistical and machine-learning approaches to DM,

another is in the relative emphasis they give to models and algorithms. Modern statistics is almost entirely driven by the notion of a model. This is a postulated structure, or an approximation to a structure, which could have led to the data. In place of the statistical emphasis on models, machine learning tends to emphasize algorithms. This is hardly surprising; the very word "learning" contains the notion of a process, an implicit algorithm. Basic modeling principles in DM also have roots in control theory, which is primarily applied to engineering systems and industrial processes. The problem of determining a mathematical model for an unknown system by observing its input-output data pairs is generally referred to as system identification. The purposes of system identification are multiple and, from a standpoint of DM, the most important are to predict a system's behavior and to explain the interaction and relationships between the variables of a system.

2.1 THE NEED OF DATA MINING AND DATA WAREHOUSING IN KNOWLEDGE MANAGEMENT

We begin with data, which is just a meaningless point in space and time, without reference to either space or time. It is like an event out of context, a letter out of context, a word out of context. The key concept here is "out of context." And, since it is out of context, it is without a meaningful relation to anything else. When we encounter a piece of data, if it gets our attention at all, our first action is usually to attempt to find a way to attribute meaning to it. We do this by associating it with other things. For example, if we see the number 6, we can immediately associate it with cardinal numbers and relate it to being greater than 5 and less than 7, whether this was implied by this particular instance or not. If we see a single word, such as "time," there is a tendency to immediately form associations with previous contexts within which we have found "time" to be meaningful. This might be, "being on time," "a stitch in time saves nine," "time never stops," etc. The implication here is that when there is no context, there is little or no meaning. So, we create context but, more often than not, that context is somewhat akin to conjecture, yet it fabricates meaning. That a collection of data is not information, as Neil indicated, implies that a collection of data for which there is no relation between the pieces of data is not information. The pieces of data may represent information, yet whether or not it is information depends on the understanding of the one perceiving the data. We would also tend to say that it depends on the knowledge of the interpreter, but we will probably getting ahead of our self, since we haven't defined knowledge. What we will say at this point is that the extent of our understanding of the collection of data is dependent on the associations we able to discern within the collection. The associations we able to discern are dependent on all the associations we have ever been able to realize in the past. Information is quite simply an understanding of the relationships between pieces of data and other information. While information entails an understanding of the relations between data, it generally does not provide a foundation

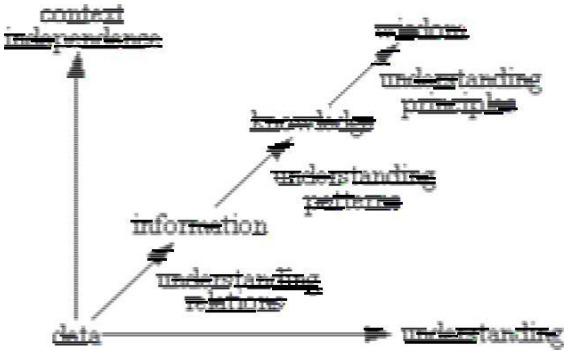

Figure 2.1

for why data is, what it is, nor an indication as to how the data is likely to change over time. Information has a tendency to be relatively static in time and linear in nature. Information is a relationship between data and, quite simply, is what it is, with great dependence on context for its meaning and with little implication for the future. Beyond relation there is pattern [bat88] [13], where pattern is more than simply a relation of relations. Pattern embodies both a consistency and completeness of relations which, to an extent, creates its own context. Pattern also serves as an Archetype [sen90] [14] with both an implied repeatability and predictability. When a pattern relation exists amidst the data and information, the pattern has the potential to represent knowledge. It only becomes knowledge, however, when one is able to realize and understand the patterns and their implications. The patterns representing knowledge

have a tendency to be more self-contextualizing. That is, the pattern tends, to a great extent, to create its own context rather than being context dependent to the same extent that information is. A pattern which represents knowledge also provides, when the pattern is understood, a high level of reliability or predictability as to how the pattern will evolve over time, for patterns are seldom static. Patterns which represent knowledge have completeness to them that information simply does not contain.

Wisdom arises when one understands the foundational principles responsible for the patterns representing knowledge being what they are. And wisdom, even more so than knowledge, tends to create its own context. We have a preference for referring to these foundational principles as eternal truths, yet we find people have a tendency to be somewhat uncomfortable with this labeling. These foundational principles are universal and completely context independent. In Summary the following associations can reasonably be made:

- Information relates to description, definition, or perspective (what, who, when, where).
- Knowledge comprises strategy, practice, method, or approach (how).
- Wisdom embodies principle, insight, moral, or archetype (why).

DM can best be described as Business Intelligence (BI) technology that has various techniques to extract comprehensible, hidden, and useful information from a population of data. DM makes it possible to discover hidden trends and patterns in large amount of data. The output of DM exercise can take the form of patterns, trends, or rules that are implicit in the data. [6]. DM is an interdisciplinary field. The type of DM to use depends on the problem to be solved, and on the type of patterns and data to be mined. Moreover, depending on the specific problem, techniques from various fields can be used, including machine learning (Mitchel, 1998; Michalski, 1998; Witten and Frank, 1999), artificial intelligence (Russel and Norvik,2002), statistics, information retrieval (Baeza-Yates et al., 1999), natural language processing (Manning and Schütze, 1999; Dale et al., 2000), pattern recognition and visualization. Different patterns can be mined using different DM functionalities such as concept/class description, association analysis, classification and regression, cluster analysis, trend analysis, deviation analysis and similarity analysis. The process of knowledge discovery generally involves an interactive sequence of the following steps: data cleaning, data integration, data selection, data transformation, modeling, pattern evaluation, and knowledge representation (Han and Kamber, 2001). During the preprocessing steps - data cleaning and integration, data is analyzed in order to remove noise and inconsistencies. The resulting preprocessed data are stored in a data warehouse. During the modeling step, intelligent techniques are used to extract and then evaluate data patterns. The patterns found relevant can be presented to users using visualization and representation techniques. DM is often used to build predictive/inference models aimed to predict future trends or behaviors based on the analysis of structured data. In this context, prediction is the construction and the use of a model to assess the class of an unlabeled example or to assess the values ranges of an attribute that a given example is likely to have. Classification and regression are the two major types of prediction problems, where classification is used to predict discrete or nominal values while regression is used to predict continuous or ordered values. Although most DM techniques focus on mining structured data, an increasingly important task in DM is to mine complex and heterogeneous types of data. In reality, a substantial portion of the available information is stored in document repositories, which consist of a large collection of documents from various sources such as research papers, books, digital libraries, e-mail messages, articles, and web pages etc.,. Data stored in most text databases are semi-structured: neither completely unstructured nor completely structured. Traditional information retrieval techniques become inadequate due to these increasingly vast and heterogeneous amounts of text data (Han and Kamber, 2001). Users need tools to compare different documents, rank the importance or the relevance of the documents, or find patterns and trends across multiple documents. Thus text mining has become an increasingly

popular and essential theme in data mining (Han and Kamber, 2001). Now a day's technology automate the mining process, integrates it with commercial data warehouses, and presents it in a relevant way for many business users.

2.2 Evaluating the Benefits of a Data Mining Model Figure 2-2, which shows a "gains chart,"

suggests some benefits available through data mining. The diagonal line illustrates the number of responses expected from a randomly selected target audience. Under this scenario, the number of responses grows linearly with the target size. The top curve represents the expected response if you allow the model scores to determine the target audience. The target is now likely to include more positive responders than in a random selection of the same size.

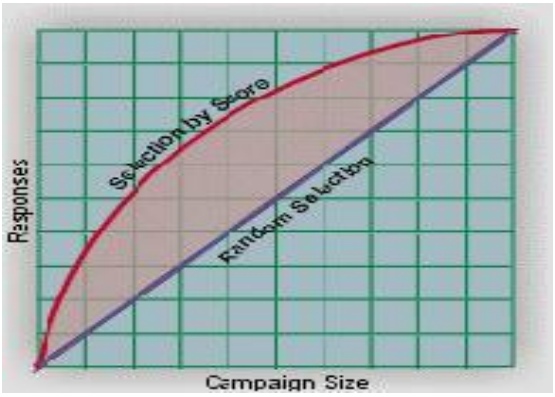

Figure2-2 Gains-Chart

The shaded area between the curve and the line indicates the quality of the model. If the curve is very steeper then better is the model. Other representations of the model often incorporate expected costs and expected revenues to provide the most important measure of model quality: profitability. A profitability graph such as Figure 2.3 can help determine the number of prospects to include in a campaign. In this example, it is easy to see that contacting all customers will result in a net loss. However, selecting a threshold score of approximately 0.8 will maximize profitability.

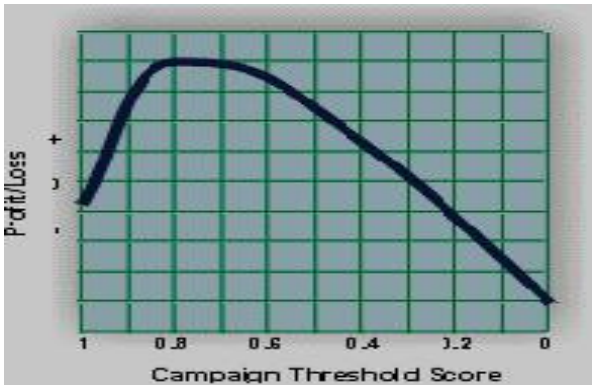

Figure2-3 Profitability-Chart

In this example, it is easy to see that contacting all customers will result in a net loss. However, selecting a threshold score of approximately 0.8 will maximize profitability.

3. DATA WAREHOUSES

In many organizations, one of the first KM tools is a Data warehouse. Corporations recognize that information placed in the hands of decision makers is a powerful tool. To meet decision makers' nearly insatiable appetite for information, data is being extracted from operational systems and placed in data warehouses. Data warehouses contain historical data organized by key business dimensions. For example, a data warehouse for a retailer contains daily product sales for each store. A data warehouse for a bank contains customer information for each bank service. Each warehouse summarizes individual transactions into time-series data for monitoring and analyzing performance Data warehouses differ from traditional transaction databases in that they are designed to support decision making rather than simply efficiently capturing transaction data. Typically, data warehouses contain many years of transaction databases stored in the same database. Data warehouses are not updated on a transaction-by-transaction basis. Instead, the entire database is updated periodically. For the maintenance and security reasons the Data warehouse is read only when compared with operational data bases. The size of data warehouses can be substantial. With all the data accessible in one place, relationships between data elements can be more effectively explored. Users can browse the data or establish queries, though this type of analysis generally results only in knowledge for particular individuals. An alternative approach is to use a process called knowledge discovery to determine whether there is additional knowledge hidden in the data. [3][8].

4. KNOWLEDGE WAREHOUSES

Knowledge warehouses are aimed more at qualitative data rather than the kind of quantitative data typical of data warehouses. KM systems generate knowledge from a wide range of databases including Lotus Notes databases, data warehouses, work processes, news articles, external databases, Web pages and people. Thus, knowledge warehouses are likely to be virtual warehouses where the knowledge is dispersed across a number of servers. In some cases, a Web browser can be used as an interface to a relational database. For example, Ford Research and Development uses a browsable Oracle database. The database contains manuals and design rules, specifications, and requirements. Another frequently used corporate application is a human resource knowledge base about employee capabilities and skills. Employee information can include name, employee-number, sex, education, specialties, previous experience, and other descriptors. Historically, Lotus Notes has provided one of the primary tools for storing qualitative and document based information and for facilitating virtual groups. With the recent explosion of the Internet, however, low-cost Web-based solutions within intranet environments have become the focus of KM. Data and knowledge bases Knowledge can come from top-down activity, work processes, news reports, and a wide range of other sources. Knowledge typically captured to meet top down requirements includes manuals, directories, and newsletters. Knowledge bases capturing information generated from work processes are likely to include working papers, proposals, and other similar documents. In addition, knowledge bases can be designed to provide continuity and history in activities like customer support. [3].

5. ONTOLOGY

Ontology is an explicit specification of a conceptualization. In enterprise KM systems, ontology specifications can refer to taxonomies of the tasks that define the knowledge for the system. Ontologies define the shared vocabulary used in the KM system to facilitate search, storage, representation and communication. Development and maintenance of an enterprise-wide ontology requires continual effort to evolve the ontology over time. Ontologies are particularly important in ensuring that best-practices databases are able to communicate to the user the broadest range of practices and activities and allow the user to recognize when a best practice would fit in their organization. Price Waterhouse reportedly has ontology with over 4,500 entries for its best-practices database. Since Price Waterhouse is an international firm, the ontology has been translated into other languages to broaden use and accessibility of the knowledge base. In addition, since enterprises are often involved in multiple industries, multiple ontology may be required as part of the KM system. Out of necessity, virtually all enterprises with a KM system have developed their own ontology. Because these firms have made this investment, ontology construction appears at this point to offer competitive advantages. However, at least one firm has expressed interest in an ontology shared across multiple organizations in order to cut development costs and to speed system development. Over time, industries are likely to form coalitions or subscribe to central services for these reasons. Other knowledge description attributes In addition to ontology information, additional descriptive attributes of the knowledge can prove critical to its use and maintenance. Contributor, organization, and status information are all viable descriptive attributes. Virtually all knowledge bases capture contact or contributor information, including contact or contributor names, date of contribution, and the person's role in generating the knowledge and so on. Many knowledge bases also include organizational information that can include the division or department in which the project was built or from which the knowledge was gathered. Status information about knowledge is also a typical kind of descriptive attribute.

6. KNOWLEDGE DISCOVERY:

Generating the knowledge from data is nothing but knowledge discovery. Knowledge discovery is a new and rapidly evolving discipline that uses tools from AI, mathematics, and statistics to tease knowledge out of data warehouses. Gregory Piatetsky-Shapiro and William Frawley define knowledge discovery as "nontrivial extraction of implicit, previously unknown, and potentially useful information from data." Because knowledge discovery approaches can be designed to exploit characteristics and structures of the underlying application domain, knowledge discovery has found use in a wide range of applications, including fraud analysis, credit card analysis, security, customer analysis, and product analysis. Knowledge discovery is a method that includes different approaches and tools to analyze both numeric data and text. For example, organizations have developed different ways to generate knowledge from numeric databases, such as the financial information in the US Security and Exchange Commission's Edgar (Electronic Data Gathering and Retrieval System). Price Waterhouse developed an intelligent system called EdgarScan,, to make Edgar available on the Web (http://edgarscan. tc.pw.com). EdgarScan lets users access a repository of publicly available financial information. Data is periodically extracted from the Edgar [3].

CONCLUSION:

In information systems, most research on knowledge management assumes that knowledge has positive implications for organizations. However, knowledge is a double-edged sword: while too little might result in expensive mistakes, too much mind, commodity, and discipline. One of the concepts for efficient managing codified knowledge is data mining. As employees turn over in today's overheated job market, organizations are likely to lose access to large quantities of critical knowledge. By using Data mining - technologies and techniques for recognizing and tracking patterns within data - helps businesses sift through layers of seemingly unrelated data for meaningful relationships, where they can anticipate, rather than simply react to, customer needs. AI techniques, DM and Data warehouses and allied concept support to create a system that will capture company-wide knowledge and make it widely available to all its members.

REFERENCES

[1] Han, J. and M. Kamber, Data Mining: Concepts and Techniques, (Morgan Kaufmann, San Francisco, 2000.)

[2] Vladimir KVASSOV and Sara C. MADEIRA "Using Data Mining Techniques for Knowledge Management: an Empirical Study"

[3] Daniel E, O'Leary, "Enterprise Knowledge Management," IEEE Computer, March, 1998, pp. 54-61.

[4] Brian R. Gaines, "The Emergence of Knowledge through Modeling and Management Processes in Societies of Adaptive Agents".

[5] Agresti, W., "Knowledge Management," Advances in Computers, Vol. 53, 2000, pp. 171- 283

[6] Berson, A., S. Smith, K. Thearling, Building Data Mining Applications for CRM (McGraw-Hill, New York, 2000.)

[7] Alex Berson, and Stephen J. Smith, "Data Warehousing, Data Mining, & OLAP" (TATA McGraw-HILL Edition 2004.)

[8]Amitesh Sinha, "Data Warehousing" (Thomson,2002.)

[9] Pieter Adriaans and Dolf Zantinge, "Data mining" (Pearson education, 2002)

[10] Alberthal, Les. Remarks to the Financial Executives Institute, October 23, 1995, Dallas, TX

[11] Bellinger, Gene. The Knowledge Centered Organization.

[12] Bateson, Gregory. Mind and Nature: A Necessary Unity, Bantam, 1988

[13] Senge, Peter. The Fifth Discipline: The Art & Practice of the Learning Organization, Doubleday-Currency, 1990.