

INTELLIGENT TECHNOLOGIES FOR KNOWLEDGE DISCOVERY

ADRIAN COJOCARIU, CRISTINA OFELIA STANCIU

"TIBISCUS" UNIVERSITY OF TIMIȘOARA, FACULTY OF ECONOMIC SCIENCE, DALIEI STR, 1/A,
TIMIȘOARA, 300558, ROMANIA

a_cojocariu@yahoo.com, ofelia.stanciu@gmail.com

Abstract:

Knowledge, as intellectual capital, has become the main resource of an organization, and the process of knowledge discovery, acquisition and storage is a very important one. Knowledge discovery can be easily realized through Data Mining, a Machine Learning technique, which allows the discovery of useful knowledge from a large amount of data, this knowledge supporting the decision process. A proper knowledge management of the discovered knowledge is able to improve the organization's results and will lead to increasing the intellectual capital, the result being a more efficient management.

Key words: knowledge, data mining, data warehouse, Machine Learning

JEL classification: M15, C80, D80

I. INTRODUCTION

Within the knowledge society, the most important asset of any organization is knowledge, as intellectual capital, and the process of knowledge discovery, acquisition and storage is a very important one.

Knowledge discovery refers to the ways of generating knowledge discovering or deriving it from existing information, which has resulted from data processing. The discovered knowledge will be organized, by indexing, cleaning and setting relationships between the elements. One of the main stages of the of the knowledge discovery process is Data Mining, which provides a series of techniques for pattern discovery within a very large data base.

Knowledge management is a process which involves knowledge organizing, sustaining, applying, sharing and improving, aiming to increase the organization's performances and assuring a higher adaptation level to the environmental requests. The improvement of knowledge management at the organization level implies an efficient data storage system, meant to lead to an intelligent knowledge discovery, by processing the resulted information.

II. THE KNOWLEDGE DISCOVERY PROCESS. DATA WAREHOUSING AND MODERN DATA PROCESSING

The process of knowledge discovery is an interactive and iterative process which assumes several stages, which are shown in Figure no.1.

Considering a large amount of data that exist within the organization, only the data which seems to be important will be selected, this data being known as target data.

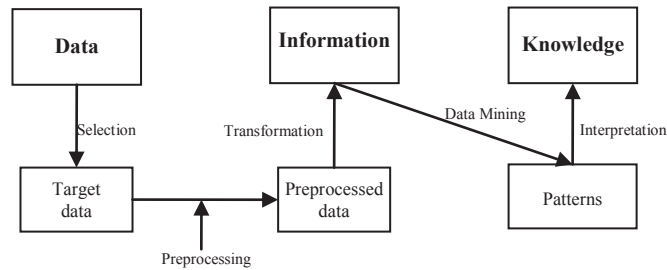


Figure no. 1. The knowledge discovery process
(Adapted by [6])

During the preprocessing stage, after the data cleaning and removing the residual data, one can establish the approach strategy of the data containing unknown fields, with unknown values. The data transformation stage is aiming to represent data in an optimal way by generating some methods to reduce the number of variables considered. A Data Mining function, such as summarization, classification, regression or clustering follows, as it is necessary the use of statistic algorithms, visualization techniques and decision trees in order to establish methods for pattern discovery. During the interpretation stage, applying reasoning methods upon the discovered patterns will lead to new discovered knowledge within the organization.

The Data Warehouse is a system that contains the operational and historical data of an organization, different from all the other operational data bases. The large amount of data contained by a Data Warehouse is provided by internal sources as well as external sources of the organization. The Data Warehouse is taking data from the operational data bases, and afterwards different analysis will be made upon the data, in order to support the decision process.

According to W.H. Inmon, Data Warehouses are “a collection of subject oriented data, integrated, historical and permanent intended to support the decision process”, from this definition deriving the Data Warehouses characteristics: subject orientation, integration, historic character and data permanency.

The Data Warehouses building and using process is known as Data Warehousing, process that implies data integration, cleaning and normalization.

The objects of a Data Warehouse are the following:

- good access assurance to data for the users – a microcomputer facilitates the users access to the Data Warehouse;
- providing a single version of reality – the user will not have to deal with ambiguous data, so there will be no debate upon the data;
- the recording and accurate restoring of the past – data from the past can be extremely important for the decision makers, because most of the time the present data is irrelevant unless it is compared with data from the past;
- allowing the combined synthesis/detail access to data – the collection and processing of information is being done much easier using data from the Data Warehouse;
- making a clear difference between operational and analytical level processing - the maintenance of an information system if the information used for decisions and the operational information is considered the same is difficult.

Data Warehouses contain different types of data: detailed data, aggregated data, metadata, the last ones, considering their source and their transformation rules, being used to records configuration within the Data Warehouse.

The main ways different categories of decision makers will benefit of the data kept in Data Warehouses are the on-line analytical processing (OLAP) and the Data Mining technologies. The OLAP technology refers to the possibility of aggregation of the data contained by a Data Warehouse, being capable to discover from the large amount of data information that is useful to support the decision process within an organization. According to some researchers, an alternative term, more representative to describe the OLAP concept, would be FASMI (Fast Analysis of Shared Multidimensional Information). The substance of each OLAP system is the OLAP cube, also known as multidimensional cube, composed by numeric facts called measures, classified by dimension [13]. These measures follow from the record of the tables from the relational data base. The results of the users' requests will be obtained by dynamic parsing of the dimensions of the data cube, at different levels of synthesis or detailing.

The characteristics of the OLAP systems are the following [9]:

- the multidimensional perspective upon the data;
- intensive computational capacity;
- time intelligence.

The technologies available for data and information management ought to contribute to a better understanding of the past and a better forecast of the future by enhancing the decision process, using Data Mining technologies.

Within the decision support systems, integrated Data Mining technologies lead to a decision support tool based on the human-machine interaction. These two entities, as a whole, represent a specter of information technologies for analysis, realizing an optimal combination for generating an intelligent system, though autonomous, still human coordinated [3].

Figure no. 2 shows the elements that must be found within the decision support system of an organization that is using high technologies. According to the organizations complexity, some of the elements may or may not be present.

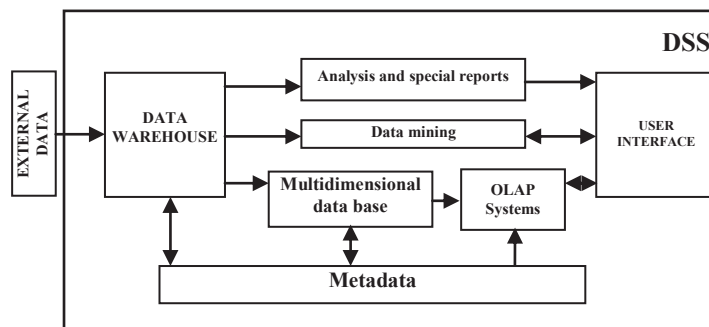


Figure no. 2. The elements within the decision support systems of an organization that is using high technologies

Data Mining technologies, due to their characteristics, are very suitable to analyze large amount of data. Data Mining is aiming to discover patterns within data sets, while other analytical technologies, such as queries, statistical analysis systems are not able to, and OLAP tools are based upon verifications, which prove to be limited.

The collecting of data that reflects an organization's activity has become vital in order to achieve competitional advantage. The medium and large companies have made investments into computer based systems that collect data and are able to manage very

large data bases. The main task these systems have to successfully fulfill is knowledge discovery that follows after the reasoning upon the information that results from the collected data.

The large amount of data is outrunning the human processing capacity, also in order the decisions to be correctly grounded, systems that are using Machine Learning technologies are required. These systems allow the discovery of patterns at the very level of unprocessed data, providing different results that can be used within the decision support systems but can also be used by the human analyst.

The Data Mining process consists in four important phases: data collecting, data preparation, pattern discovery and pattern analysis. Data collecting phase involves overtaking data from different sources, and considering that this data could be heterogeneous, the preparation phase will normalize the data and represent it in structures, in order to facilitate the data use. The data identified after certain characteristics during the previous phase is extracted and afterwards formatted, so the data will be represented in the form that the Data Mining application requires. The discovery of new patterns follows from applying Data Mining technologies upon the selected data.

Nowadays, computer based information systems that are using Data Mining technologies are able “to learn” from the previous behavior of the considered elements, and based on the knowledge following from the “learning process” they are able to make hypothesis which they will be testing. The knowledge that proves to be valid and useful can be integrated within the decision support systems, in order to be useful to the decision makers and assist them in making the right decision.

The high necessity, popularity and utility of Data Mining technologies is increasing due to the following reasons:

- the increase of the amount of daily gathered data within an organization can be overwhelming;
- the limitation of human analysts, whose analysis capacity is overtaken by the large amount of data and the complexity that is caused by the data dependencies. Besides, human analysis may bring a deficit of objectivity;
- the low costs of automated Data Mining systems, comparing to the costs that a larger number of human specialists would imply. Though, Data Mining doesn't totally eliminate the human involvement, but it facilitates it tasks, allowing to an analyst that is no specialist in statistics or programming to discover the required knowledge.

Data Mining technologies can accomplish the following tasks:

- prediction – future values of variables we care about can be acquired by finding patterns in examples and developing a model;
- classification – finding a function that classifies the records into discreet classes;
- relation detection – allows searching for the most influent independent variables;
- explicit modeling – describing different variable dependencies through explicit formulas;
- clustering – allows identifying similar record groups that are different from other records not in the group. It is often needed to also identify the variables that lead to obtaining the best clusters.

Patterns established by Data Mining technologies that are generated using prediction techniques prove themselves to be highly important throughout the decision-making process because they bring to light various aspects that can lead to an improvement of the decision-making process, from an efficiency point of view as well as from a time-consuming one.

Regarding software systems designed for Data Mining, these have been divided in two categories: Data Mining tools and Data Mining applications. The Data Mining tools offer techniques that apply in any situation and for every given problem. Data Mining applications are the ones “encapsulating” Data Mining techniques within an applications designed for solving a specific problem. This separation does not imply that one of the categories is less important and requires less of the organizations' attention.

Data Mining tools must ensure some level of flexibility and a great accuracy in order to improve the efficiency of Data Mining applications. Given the fact that existing data differ from one organization to another, it is difficult to develop a technique that would help acquire the optimum results for any given organization. Data Mining tools are flexible from this point of view and allow combining of techniques in order to improve the accuracy of the desired outcome.

Data Mining technologies based on Machine Learning algorithms represent the means to discover relations (*patterns*) in the input collections, using them to develop new rules and to predict future behavior, succeeding in transforming the enormous amount of data into valuable knowledge that leads to many business opportunities.

III. KNOWLEDGE MANAGEMENT SYSTEMS

In the 90's a new concept of leadership has developed within organizations, called *knowledge management*, which is based on the idea of intellectual capital to identify knowledge as untouchable assets.

The purpose of knowledge management is to ensure the distribution and development of knowledge within an organization. Knowledge is created, shared and spread to the members of the organization based on implicit and explicit rules, thus causing an increase in the knowledge of those members; this process is called *organizational learning*. Organizational learning implies innovation that leads to a continuous improvement of the whole organization, given the conditions of the harsh competition in the business and organizational environment.

One of the primary objectives of knowledge management is using technologies in order to ease generating, accessing and reusing knowledge.

The activities within knowledge management take place on several levels [1]:

- a strategic level which is attempting to find, transfer and spread knowledge that refer to the long term goals of the organization;
- a tactical level which is leaning towards finding, transferring and spreading knowledge on a medium term basis;
- an operational level which is related to the short term operations of the organization.

Any of these levels though requires the knowledge to be defined and a finding a way to transfer knowledge to systems or human users.

The knowledge management system is described as being a ever-evolving system, its inner processes being depicted in Figure no. 3.

A knowledge management example is the document management system within an organization, which offers control upon the knowledge gathered from the information within those documents. Information within an organization must be shared to some persons which prove to be interested in it. Due to the fact that information is usually found within different types of documents, developing a database to store, index and secure each document is in order. Maintaining and managing physical documents implies manually accessing the information, which is a time-consuming process even if a well-developed archiving system exists. Moreover, such storage manner can easily

bring damage to the documents and thus bringing a negative effect to the organization, from a strategic, legal and income-related point of view.

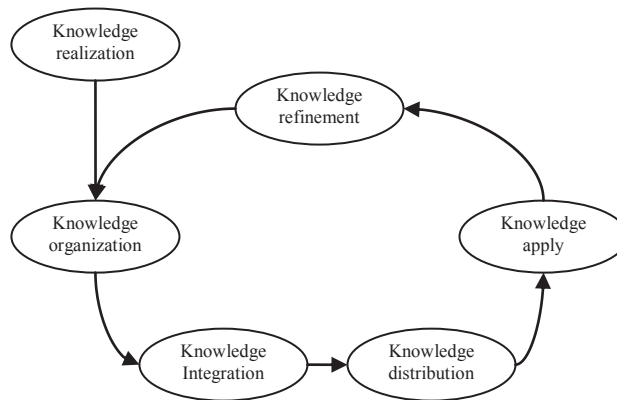


Figure no.3. The knowledge management process
(Source:[6])

A computer-based system for document management would have to be very well organized, to present a friendly user interface and to embed a fast and secured communication system for within the organization as well as for the outside world, for communicating to other organizations providing they own an Internet connection. The functionality of such a system should also ensure recording and distributing documents at the highest level of security standards, should monitor the status of documents as so only authorized persons can retrieve the status of any document at some point, should be able to perform and present various statistical analysis upon the documents. These document management systems represent the best way to manage knowledge within an organization.

IV. CONCLUSION

Within the knowledge discovery process Data Mining technologies prove to be very useful, offering a series of ways to discover patterns in very large data bases. We have shown that the knowledge discovery process is an interactive and iterative process that assumes several stages, starting from the raw data. An organization's data is stored within a Data Warehouse that is taking data from the operational data bases, and afterwards different analysis will be made upon the data in order to support the decision process. The two main ways that different categories of decision makers can benefit of the data from Data Warehouses have been presented within the paper. The two ways are the on-line analytical processing and Data Mining technologies. The Data Mining technologies prove to be great tools for the analysis of very large amount of data, aiming to discover patterns within data sets, while other technologies, such as queries, stochastic analysis technologies, OLAP tools, are based on verifications, which proved to be limited.

The knowledge discovered through the techniques that were mentioned above must be carefully managed, as they represent the most valuable asset of the organization. Knowledge management is aiming to assure the distribution and improvement of the organization's knowledge. Knowledge is created, shared and distributed to the organization's members on the basis of explicit and implicit rules.

BIBLIOGRAPHY

- [1] Ackoff, R. L., *From Data to Wisdom*, Journal of Applied Systems Analysis, Volume 16, 1989, pag. 3-9
- [2] Cortes, colab., *Knowledge Management in Environmental Decision Support Systems*, AI Communications, 2001
- [3] Drăgănescu, M., *De la societatea informațională la societatea cunoașterii*, Editura Tehnică, București, 2003
- [4] Ganguly, A. R., Gupta A., *Data Mining Technologies and decision Support Systems for Business and Scientific Applications*, Encyclopedia of Data Warehousing and Mining, Blackwell Publishing, 2005
- [5] Graz, P., Watson, H., *Decision Support in the Data Warehouse*, Prentice Hall, Upper Saddle River Publishing, 1998
- [6] Hackathorn, R., Keen, P., *Organization Strategies for Personal Computing in Decision Support Systems*, MIS Quarterly, Vol. 5, No.3, 1981
- [7] Holsappale, C., Whinston, A., *Decision Support Systems: A Knowledge Based Approach*, St. Paul, West Publishing, 1996
- [8] Nițchi, Șt., Rodica Avram-Nițchi, *Data mining, o nouă eră în informatică*, Revista Byte, România, Februarie 1997, <http://www.byte.ro/byte97-02/18tend.html>
- [9] Power, D., *Categorizing Decision Support Systems: A Multidimensional Approach*, in volume *Decision Making Support Systems: Achievements, Trends and Challenges for New Decade*, Idea Group Publishing, 2003
- [10] Power, D., *Decision Support Systems Hyperbook.*, Cedar Falls, IA: DSSResources.COM, HTML version, Fall 2000, accessed on 13.11.2006 at URL <http://dssresources.com/dssbook/>.
- [11] Soundarajan, colab, *Knowledge Discovery Tools and Techniques*
- [12] Witten, I., Frank, E., *Data Mining – Practical Machine Learning Tools and Techniques, Second Edition*, Norgan Kaufmann Publishing, 2005
- [13] Zaharie D., Albulescu F., Bojan I., Ivacenco V., Vasilescu C., *Sisteme informatice pentru asistarea deciziei*, Editura Dual Tech, 2001
- [14] <http://www.machinelearning.net/>
- [15] <http://www.altimate.ca>
- [16] <http://www.crisp-dm.org/>
- [17] <http://www.wikipedia.org/>