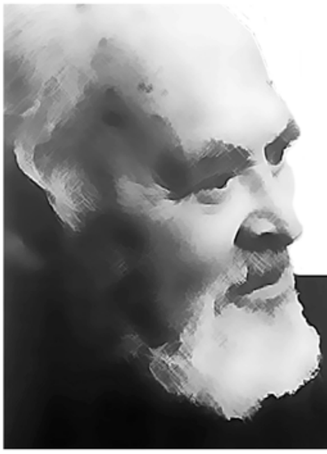

Ramón García Martínez, Paola Britos, Darío Rodríguez

Information mining processes based on intelligent systems

Preimpresión entregado por los autores al Repositorio Digital, publicado con posterioridad en *Recent Trends in Applied Artificial Intelligence. Lecture Notes in Computer Science*. Volume 7906, 2013, p. 402-410

Information Mining Processes Based on Intelligent Systems

Ramon García-Martínez, Paola Britos, Dario Rodríguez

Information Systems Research Group. National University of Lanus. Argentina.
Information Mining Research Group. National University of Rio Negro at El Bolsón. Argentina

rgarcia@unla.edu.ar, pbritos@unrn.edu.ar

Abstract. Business Intelligence offers an interdisciplinary approach (within which is Information Systems), that taking all available information resources and using of analytical and synthesis tools with the ability to transform information into knowledge, focuses on generating knowledge that contributes to the management decision-making and generation of strategic plans in organizations. Information Mining is the sub-discipline of information systems which supports business intelligence tools to transform information into knowledge. It has defined as the search for interesting patterns and important regularities in large bodies of information. We address the need to identify information mining processes to obtain knowledge from available information. When information mining processes are defined, we may decide which data mining algorithms will support the information mining processes. In this context, this paper proposes a characterization of the information mining process related to the following business intelligence problems: discovery of rules of behavior, discovery of groups, discovery of significant attributes, discovering rules of group membership and weight of rules of behavior or rules of group memberships.

1. Introduction

Business Intelligence offers an interdisciplinary approach (within which are included the Information Systems), that takes all the available information resources and the usage of analytical and synthesis tools with the ability to transform information into knowledge, focuses on generating knowledge that supports the management decision-making and generation of strategic plans at organizations [1].

Information Mining is the sub-discipline of information systems which provides to the Business Intelligence [2] the tools to transform information into knowledge [3]. It has been defined as the discovery of interesting patterns and important regularities in large information bases [4]. When speaking of information mining based on intelligent systems [5], this refers especially in the application of intelligent systems-based methods to discover and enumerate the existing patters in the information. Intelligent systems-based methods [6] allow retrieving results about the analysis of information bases that the conventional methods fail to achieve [7], such as: TDIDT algorithms (Top Down Induction Decision Trees), self-organizing maps (SOM) and Bayesian networks. TDIDT algorithms allow the development of symbolic descriptions of the data to distinguish between different classes [8]. Self-organizing maps can be applied in the construction of information clusters. They have the advantage of being tolerant to noise and the ability to extend the generalization when needing the manipulation of new data [9]. Bayesian networks can be applied to identify discriminative attributes in large information bases and detect behavior patterns in the analysis of temporal series [10].

It has been noted the necessity of having processes [11] that allow obtaining knowledge [12] from the large information-bases available [13], its characterization [14] and involved technologies [15].

In this context, this paper proposes a characterization of the information mining process related to the following business intelligence problems: discovery of behavior rules, discovery of groups, discovery of significant attributes, discovery of group-membership rules and weighting of behavior or group-membership rules, and the identification of information-systems technologies that can be used for the characterized processes.

2. Proposed Techniques for Information Mining Processes

In this section, the following information-mining processes are proposed: discovery of behavior rules (Section 2.1), discovery of groups (Section 2.2), discovery of significant attributes (Section 2.3), discovery of group-membership rules (Section 2.4) and weighting of behavior or group-membership rules (Section 2.5).

2.1. Process of Discovery of Behavior Rules

The process for discovery of behavioral rules applies when it is necessary to identify which are the conditions to get a specific outcome in the problem domain. The following problems are examples among others that require this process: identification of the characteristics for the most visited commercial office by customers, identification of the factors that increase the sales of a specific product, definition of the characteristics or traits of customers with high degree of brand loyalty, definition of demographic and psychographic attributes that distinguish the visitors to a website.

For the discovery of behavioral rules from classes attributes in a problem domain that represents the available information base, it is proposed the usage of TDIDT induction algorithms [16] to discover the rules of behavior for each class attribute. This process and its products can be seen graphically in Figure 1.

First, all sources of information (databases, files, others) are identified, and then they are integrated together as a single source of information which will be called integrated data base. Based on the integrated data base, the class attribute is selected (attribute A in the Figure).

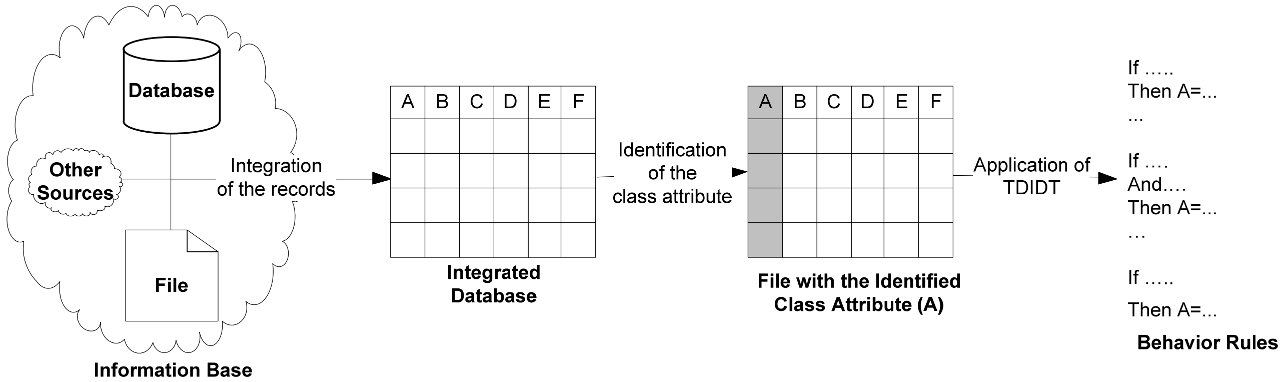

Fig. 1. Schema and products resulting of applying Process of Discovery of Behavior Rules using TDIDT

As a result of running the Process of Discovery of Behavior Rules, applying TDIDT to the class attribute, a set of rules which define the behavior of that class is achieved.

2.2. Process of Discovery of Groups

The process of discovery of groups applies when it is necessary to identify a partition on the available information base of the problem domain. The following problems are examples among others that require this process: identification of the customers segments for banks and financial institutions, identification of type of calls of customer in telecommunications companies, identification of social groups with the same characteristics, identification of students groups with homogeneous characteristics.

For the discovery of groups [17] [18] in information bases of the problem domain for which there is no available "a priori” criteria for grouping, it is proposed the usage of Kohonen’s Self-Organizing Maps or SOM [19] [20] [21]. The use of this technology intends to find if there is any group that allows the generation of a representative partition for the problem domain which can be defined from available information bases. This process and its products can be seen graphically in Figure 2.

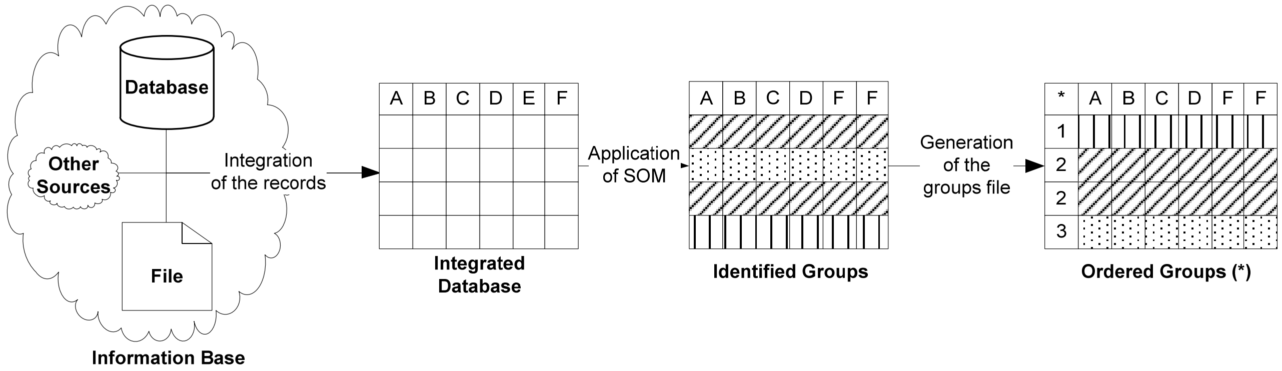

Fig. 2. Schema and products resulting of applying Process Discovery of Groups using SOM

First, all sources of information (databases, files, others) are identified, and then they are integrated together as a single source of information which will be called integrated data base. Based on the integrated data base, the self-organizing map (SOM) is applied. As a result of the application of Process Discovery of Groups using SOM, a partition of the set of records in different groups, that will be called identified groups, is achieved. For each identified group, the corresponding data file will be generated.

2.3. Process of Discovery of Significant Attributes

The process of discovery of significant attributes applies when it is necessary to identify which are the factors with the highest incidence (or occurrence frequency) for a certain outcome of the problem. The following problems are examples among others that require this process: factors with incidence on the sales, distinctive features of customers with high degree of brand loyalty, key-attributes that caracterize a product as marketable, key-features of visitors to a website.

Bayesian Networks [22] allows to see how variation in the values of attributes, impact on the variation of the value of class attribute. The use of this pocess seeks to identify whether there is any interdependence among the attributes that modelize the problem domain which is represented by the available information base. This process and its products can be seen graphically in Figure 3.

First, all sources of information (databases, files, others) are identified, and then they are integrated together as a single source of information which will be called integrated data base. Based on the integrated data base, the class attribute is selected (attribute A in the Figure 3).

As a result of the application of the Bayesian Networks structural learning to the file with the identified class attribute, the learning tree is achieved. The Bayesian Networks predictive learning is applied to this tree obtaining the tree of weighting interdependence which has the class attribute as a root and to the other attributes with frequency (incidence) related the class attribute as leaf nodes.

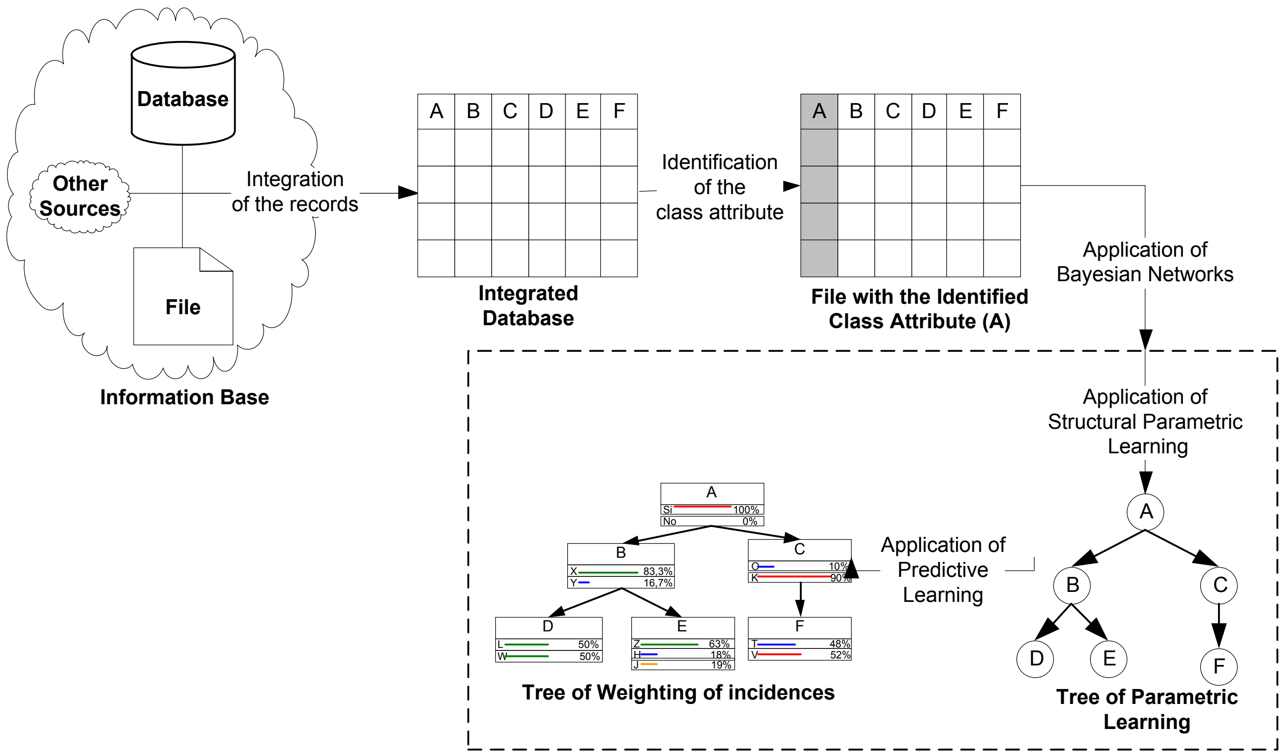
Database
Other Sources
File
Information Base
Integration of the records
A B C D E F
Integrated Database
Identification of the class attribute
A B C D E F
File with the Identified Class Attribute (A)
Application of Bayesian Networks
Application of Structural Parametric Learning
A
B
C
D
E
F
Tree of Parametric Learning
Application of Predictive Learning
A
Si 100%
No 0%
B
X 83,3%
Y 16,7%
C
O 10%
K 90%
D
L 50%
W 50%
E
Z 63%
H 18%
J 19%
F
T 48%
V 52%
Tree of Weighting of incidences

Fig. 3. Schema and resulting products for applying Process Discovery of Significant Attributes using Bayesian Networks

2.4. Process of Discovery of Group-membership Rules

The process of discovery of group membership rules applies when it is necessary to identify which are the conditions of membership to each of the classes of an unknown partition “a priori”, but existing in the available information bases of the problem domain.

The following problems are examples among others that require this process: types of customer’s profiles and the characterization of each type, distribution and structure of data of a web site, segmentation by age of students and the behavior of each segment, classes of telephone calls in a region and the characterization of each class.

For running the process of discovery of group-membership rules it is proposed to use of self-organizing maps (SOM) for finding groups and; once the groups are identified, the usage of induction algorithms (TDIDT) for defining each group behavior rules [23] [24] [21]. This process and its products can be seen graphically in Figure 4.

2.5. Process of Weighting of Behavior or Group-membership Rules

First, all sources of information (databases, files, others) are identified, and then they are integrated together as a single source of information which will be called integrated data base. Based on the integrated data base, the self-organizing maps (SOM) are applied. As a result of the application of SOM, a partition of the set of records in different groups is achieved which is called identified groups. The associated files for each identified group are generated. This set of files is called "ordered groups". The "group" attribute of each ordered group is identified as the class attribute of that group, establishing it in a file with the identified class attribute (GR). Then is applied TDIDT to the class attribute of each "GR group" and the set of rules that define the behavior of each group is achieved.

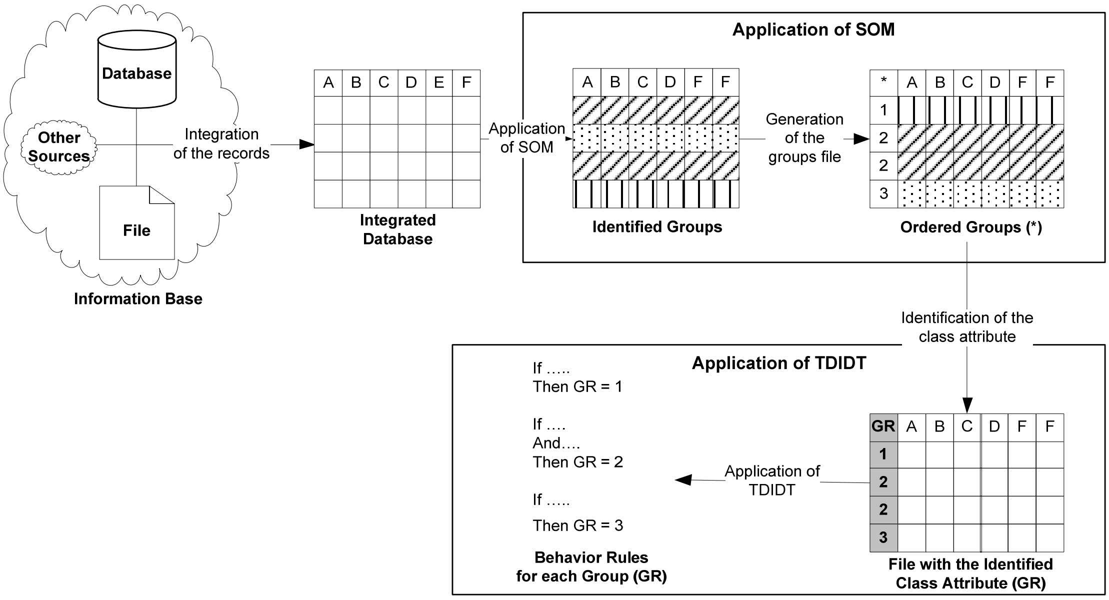

Fig. 4. Schema and resulting products of running Process of Discovery of Group-membership Rules using SOM and TDIDT

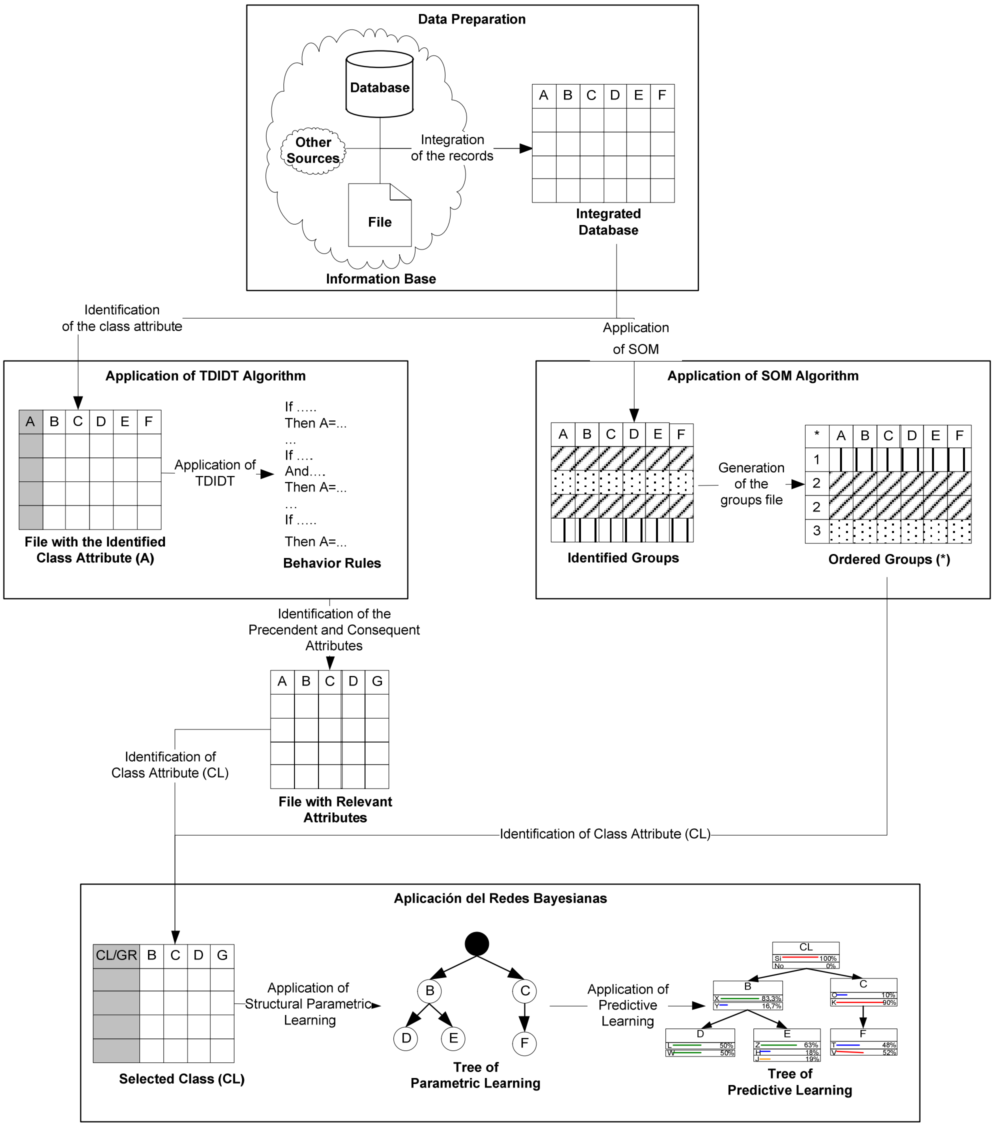
Data Preparation
Database
Other Sources
File
Information Base
Integration of the records
A B C D E F
Integrated Database
Identification of the class attribute
Application of SOM
Application of TDIDT Algorithm
A B C D E F
File with the Identified Class Attribute (A)
Application of TDIDT
If …..
Then A=...
...
If ….
And….
Then A=...
...
If …..
Then A=...
Behavior Rules
Application of SOM Algorithm
A B C D E F
Identified Groups
Generation of the groups file
* A B C D E F
1
2
2
3
Ordered Groups (*)
Identification of the Precendent and Consequent Attributes
A B C D G
File with Relevant Attributes
Identification of Class Attribute (CL)
Identification of Class Attribute (CL)
Aplicación del Redes Bayesianas
CL/GR B C D G
Selected Class (CL)
Application of Structural Parametric Learning
B
C
D
E
F
Tree of Parametric Learning
Application of Predictive Learning
CL
Si 100%
No 0%
B
X 83,3%
Y 16,7%
C
O 10%
K 90%
D
L 50%
W 50%
E
Z 63%
H 18%
J 19%
F
T 48%
V 52%
Tree of Predictive Learning

Fig. 5. Schema and resulting products of running process of weighting of behavior or Group-membership rules using SOM, TDIDT and Bayesian Neworkts

The procedure to be applied when there are classes/groups no identified includes the identification of all sources of information (databases, files, others), and then they are integrated together as a single source of information which will be called integrated data base. Based on the integrated data base, the self-organizing maps (SOM) are applied. As a result of the application of SOM, a partition of the set of records in different groups is achieved. These groups are called identified groups. For each identified group, the corresponding data file will be generated. This set of files is called "ordered groups". The group attribute of each "ordered group" is identified as the class attribute of that group, establishing it in a file with the identified class attribute (GR). As a result of the application of the structural learning, the learning tree is achieved. The predictive learning is applied to this tree obtaining the tree of weighting interdependence. The root is the group attribute and the other attributes as leaf nodes labeled with the frequency (incidence) on the group attribute.

3. Validation of the Proposed Information Mining Processes

The proposed information mining processes have been validated in three domains: political alliances, medical diagnosis and user behavior. A full detailed report of these validations can be seen in [26].

In the political alliances domain, it has been sought to discover the behavior of democrats and republicans representatives of the U.S. Congress based on the political agenda of a regular session, identifying the intraparty and interparty agreements and disagreements between interparty groups and minority intraparty groups. The first one was obtained using the process of discovery of behavior rules for the representatives of each party, and the second one by using the process of discovery of the groups for the representatives who voted homogeneously (regardless of their party affiliation) and the rules that define that homogeneity (rules of membership to each group). Additionally it has been tried to identify which have been the law or laws with greater agreement among the identified agreements, using the process of weighting of behavior rules or group-membership rules.

In the medical diagnosis domain, it has been synthesize the knowledge that allows to diagnose the type of lymphoma using as input the characteristics observed in the associated lymphography, identify the significant characteristic related to each type of diagnosis and whether there are common characteristics to different types of pathologies. The first one was obtained using the process of discovery of behavior rules for each type of diagnosis, the second one by using the process of weighting of behavior rules and the third one by using the discovery of lymphoma groups with homogeneous characteristics (regardless of its type) and the rules that defines the homogeneity (rules of membership to each group).

In the user behavior domain, it has been sought to specify a description of the reasons for subscribing or unsubscribing to a Internet service “dial-up” provided by a telephone company and identify the reasons with the highest incidence in each behavior. The first one was sought using the process of discovery of behavior rules for subscribing or unsubscribing to the service, and the second one by using the process of weighting of behavior rules.

4. Conclusions

In this paper it has been proposed and described five information mining processes: discovery of behavior rules, discovery of groups, discovery of significant attributes, discovery of group-membership rules and weighting of significant atribute related to behavior or membership rules.

Each process has been associated with the following techniques: the usage of TDIDT algorithm applied to the discovery of behavior rules or group-membership rules, the usage of self-organizing maps applied to the discovery of groups, the usage of Bayesian networks applied to the weighting of interdependence between attributes, the usage of self-organizing maps and TDIDT algorithms applied to the discovery of group-membership rules and the usage of Bayesian networks applied to the weighting of significant atribute in behavior or group-membership rules.

During the documental research work it has been noted the indiscriminate use of terms "data mining" and "information mining" to refer to the same body of knowledge. However, raising this equivalence is similar to say that computer-systems are equivalent to information-systems. The first ones are related to the technology that supports the second ones and this is what makes them different.

In this context is an open problem the need of organizing the body of knowledge related to engineering of information mining, establishing that data mining is related to algorithms; and information mining is related to processes and methologies.

On the other hand, there are in the literature many papers and results about the convenience of the usage of certain data mining algorithms compared to others, but it is rarely raised the information mining process associated to these algorithms or the convenience of the usage of one algorithm compared to other for that process. In this context, is an interesting open problem the identification of the relationship between the data mining algorithm and the process of information mining.

5. References

1. Thomsen, E. (2003). *BI's Promised Land*. Intelligent Enterprise, 6(4): 21-25.
2. Negash, S., Gray, P. (2008). *Business Intelligence*. En Handbook on Decision Support Systems 2, ed. F. Burstein y C. Holsapple (Heidelberg, Springer), Pág. 175-193.
3. Langseth, J., Vivatrat, N. (2003). *Why Proactive Business Intelligence is a Hallmark of the Real-Time Enterprise: Outward Bound*. Intelligent Enterprise 5(18): 34-41.
4. Grigori, D., Casati, F., Castellanos, M., Dayal, u., Sayal, M., Shan, M. (2004). *Business Process Intelligence*. Computers in Industry 53(3): 321-343.
5. Michalski, R. Bratko, I. Kubat, M. (1998). *Machine Learning and Data Mining, Methods and Applications* (Editores) John Wiley & Sons.
6. Kononenko, I. y Cestnik, B. (1986). *Lymphography Data Set*. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml/datasets/Lymphography. Último acceso 29 de Abril del 2008.
7. Michalski, R. (1983). *A Theory and Methodology of Inductive Learning*. Artificial Intelligence, 20: 111-161.
8. Quinlan, J. (1990). *Learning Logic Definitions from Relations.* Machine Learning, 5:239-266
9. Kohonen, T. (1995). *Self-Organizing Maps*. Springer Verlag Publishers.
10. Heckerman, D., Chickering, M., Geiger, D. (1995). *Learning bayesian networks, the combination of knowledge and statistical data*. Machine learning 20: 197-243.
11. Chen, M., Han, J., Yu, P. (1996). *Data Mining: An Overview from a Database Perspective.* IEEE Transactions on Knowledge and Data Engineering, 8(6): 866-883.
12. Chung, W., Chen, H., Nunamaker, J. (2005). *A Visual Framework for Knowledge Discovery on the Web: An Empirical Study of Business Intelligence Exploration*. Journal of Management Information Systems, 21(4): 57-84.
13. Chau, M., Shiu, B., Chan, I., Chen, H. (2007). *Redips: Backlink Search and Analysis on the Web for Business Intelligence Analysis*. Journal of the American Society for Information Science and Technology, 58(3): 351-365.

14. Golfarelli, M., Rizzi, S., Cella, L. (2004). *Beyond data warehousing: what's next in business intelligence?***.** Proceedings 7th ACM international workshop on Data warehousing and OLAP. Pág. 1-6.

15. Koubarakis, M., Plexousakis, D. (2000). A Formal Model for Business Process Modeling and Design. Lecture Notes in Computer Science, 1789: 142-156.

16. Britos, P., Jiménez Rey, E., García-Martínez, E. (2008). *Work in Progress: Programming Misunderstandings Discovering Process Based On Intelligent Data Mining Tools*. Proceedings 38th ASEE/IEEE Frontiers in Education Conference, en prensa.

17. Kaufmann, L. y Rousseeuw, P. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons Publishers.

18. Grabmeier, J., Rudolph, A. (2002). *Techniques of Cluster Algorithms in Data Mining*. Data Mining and Knowledge Discovery, 6(4): 303-360.

19. Ferrero, G., Britos, P., García-Martínez, R., (2006). *Detection of Breast Lesions in Medical Digital Imaging Using Neural Networks*. In IFIP International Federation for Information Processing, Volume 218, Professional Practice in Artificial Intelligence, eds. J. Debenham, (Boston: Springer), Pág. 1-10.

20. Britos, P., Cataldi, Z., Sierra, E., García-Martínez, R. (2008). *Pedagogical Protocols Selection Automatic Assistance*. Notes in Artificial Intelligence 5027: 331-336.

21. Britos, P., Grosser, H., Rodríguez, D., García-Martínez, R. (2008). *Detecting Unusual Changes of Users Consumption*. In Artificial Intelligence in Theory and Practice II, ed. M. Bramer, (Boston: Springer), en prensa.

22. Britos, P., Felgaer, P., García-Martínez, R. (2008). *Bayesian Networks Optimization Based on Induction Learning Techniques*. In Artificial Intelligence in Theory and Practice II, ed. M. Bramer, (Boston: Springer).

23. Britos, P., Abasolo, M., García-Martínez, R. y Perales, F. (2005). *Identification of MPEG-4 Patterns in Human Faces Using Data Mining Techniques*. Proceedings 13th International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision'2005. Páginas 9-10.

24. Cogliati, M., Britos, P., García-Martínez, R. (2006a). *Patterns in Temporal Series of Meteorological Variables Using SOM & TDIDT*. In IFIP International Federation for Information Processing, Volume 217, Artificial Intelligence in Theory and Practice, ed. M. Bramer, (Boston: Springer), Pág. 305-314.

25. Britos, P., Dieste, O., García-Martínez, R. (2008b). *Requirements Elicitation in Data Mining for Business Intelligence Projects*. In Advances in Information Systems Research, Education, and Practice eds. George Kasper e Isabel Ramos (Boston: Springer), en prensa.

26. Britos, P. (2008). *Processes of Information Mining based on Intelligent Systems* (in spanish). PhD thesis in Computer Science. School of Computing. Universidad Nacional de La Plata. http://postgrado.info.unlp.edu.ar/Carrera/Doctorado/Tesis/Britos-Tesis%20