Paper from: *Data Mining III*, A Zanasi, CA Brebbia, NFF Ebecken & P Melli (Editors).
ISBN 1-85312-925-9

Text mining: crossing the chasm between the academy and the industry

E. M. Silva[1,2], H. A. do Prado[1,3], E. Ferneda[1]
[1] *Graduate Program on Knowledge and TI Management*
Universidade Católica de Brasília, Brazil
[2] *Radiobrás – Brazilian Government Agency – "Brazil Agency", Brazil*
[3] *Center for Agricultural Research on Savannah - Embrapa, Brazil*

Abstract

The existence of a chasm between the development phase and the adoption of new technologies has been widely recognized. Some reasons that make hard the transition academy-industry for new technology are: *(a)* the weak usability commonly presented by emergent technology in regard to the required ease of ordinary users; *(b)* few successful experiences reported; and *(c)* the lack of an adequate methodology to new tools. In this paper we argue that text mining technology is exactly in the chasm point and study the hypothesis *(c)* mentioned above. The start point of our argumentation is the contradiction posed by the extraordinary amount of information in text form - about 80% of all existing information in a company - while the amount of text mining/web mining applications does not go beyond 7%. At the same time, we observe that the available technological alternatives present an excellent level of maturity, with many functions and adequate interfaces for the common user. The research was carried out by means of a case study in which we used texts issued by a journalistic agency. In order to explore our hypothesis, we applied the CRISP-DM method that was originally conceived for data mining. The contribution of this work includes the examination of the methodological hypothesis for the lack of text mining applications, an experience report in which we describe the steps carried out to apply CRISP-DM to text mining, and the findings in the target domain.

1 Introduction

Since the early nineties, researchers in Knowledge Discovery from Databases (KDD) have dedicated intensive efforts to extract human understandable patterns from structured databases, as well as to make the whole work as automatic as

Paper from: *Data Mining III*, A Zanasi, CA Brebbia, NFF Ebecken & P Melli (Editors).
ISBN 1-85312-925-9

possible. In this way important advances have been achieved, allowing technology to cross the usual gap that occurs when results move from academy to industry. However, just recently the counterpart of structured data, pure or marked text, has received attention as a crucial source of knowledge to improve business management. In this sense, studies on clustering applied to extract meaning from huge amounts of text have been carried out. This paper departs from the reasonable question about why Knowledge Discovery from Text (KDT) has not crossed the same gap. It is really hard to understand this fact if one considers the current state-of-art in KDT, which allows the organizations to take advantage from knowledge hidden in many textual sources. In this work we apply the well-known CRISP-DM methodology in the texts issued by a Brazilian journalistic organization aiming to figure out the degree in which the company has accomplished its objectives. During the case study, we observed how a data mining method (CRISP-DM) could be fitted to the case of textual data.

2 Motivation

According to Tan [5], 80% of a company's information is contained in text documents. In contrast, a poll from Kdnuggets® [4] found that only 2% percent of all knowledge discovery applications are carried out on text databases. If we add the *web mining* applications, that use marked text, that percent goes to 7%. It is the case to ask why, in an economy of increasing competition, the advantage brought by knowledge discovery from text is not as common as one could expect.

In his popular model (Figure 1) to explain the phases of technology adoption, Moore (*op. cit. in* [1]) discusses the existence of a chasm between the ´early adopters` and the ´early majority pragmatists` that technology has to cross in order to become widely applied. The motivation for this study is our belief that text mining is exactly stalled in this chasm.

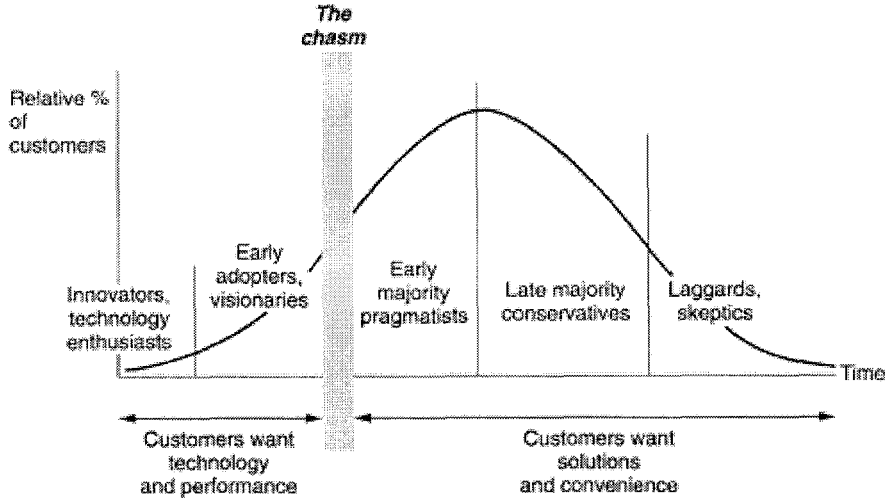

Figure 1 – Classification scheme for adopters of innovative technology

Paper from: *Data Mining III*, A Zanasi, CA Brebbia, NFF Ebecken & P Melli (Editors).
ISBN 1-85312-925-9

From this belief, we could enumerate some hypothesis to investigate why text mining has not crossed the chasm yet: *(a)* lack of adequate technology, failing, say, in usability requirements for example, *(b)* few successful experiences reported, and *(c)* lack of adequate methodology to drive users in developing text mining applications. To develop our study we focused in the third hypothesis, looking for methodological reasons for the low use of text mining technology.

3 Applying CRISP-DM

CRISP–DM (CRoss–Industry Standard Process for Data Mining) [2], is a methodology developed to promote the standardization of the data mining process. It encompasses a set of phases and processes that describe the tasks that one has to carry out to develop a data mining application. The method is vendor neutral and domain independent, being well suited to manage the whole process of development. Six phases integrate the method as shown in Figure 2 and described next.

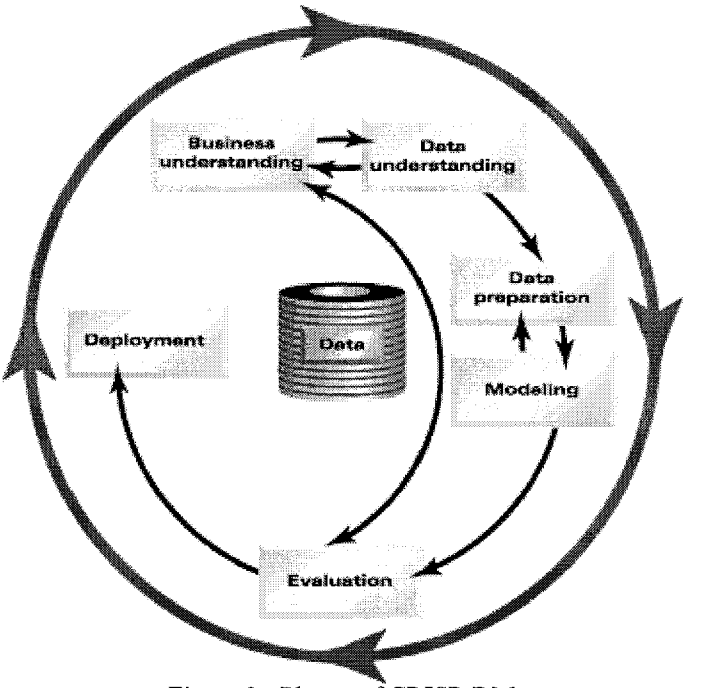

Figure 2 - Phases of CRISP-DM

Business understanding – this phase looks for the identification of requirements and objectives of the application under the client's point of view. Problems and restrictions that can cause loss of time and effort must be considered. This phase also includes a description of the client background, its business objectives, and a description of the criteria used to measure the success of the achievement.
Data understanding – identify all information relevant to carry out the study and a first approximation of its content, quality, and utility. The initial collection of data helps the analyst in learning about its details. Conflicts related to the ex-

Paper from: *Data Mining III*, A Zanasi, CA Brebbia, NFF Ebecken & P Melli (Editors).
ISBN 1-85312-925-9

pected and the real format and values are identified in this phase. Information of the manner in which data was collected, including its sources, meaning, volumes, reading procedure, etc - can also be of interest since it is a good indicator of the data quality. In this phase the first discoveries are carried out.

Data preparation – this phase consists of the tasks concerned on the acquisition of a final data set, from which the model will be created and validated. Tools for data extraction, cleaning, and transformation are applied to data preparation. Joins of tables, aggregation of values, format changing are performed to satisfy the input requirements of the learning algorithms.

Modeling – in this phase the more appropriate data mining techniques are selected and applied, according to the objectives so far defined. Modeling represents the core phase of data mining, that is, the choice of the technique, its parameterization, and its execution over a training data set. Many different and complimentary models can be created in this phase.

Evaluation – the evaluation phase consists in reviewing the past steps in order to check the results against the objectives defined in the business understanding phase. It is also defined in this phase the next tasks to be performed. According to the results, it is defined route corrections, which correspond to the return to one of the already performed phases using other parameters or looking for more data.

Deployment – set of actions necessary to make available to the organization the acquired knowledge. In this phase it is generated a final report to explain the results and the experiences useful in the client business.

3.1 Business understanding

Radiobrás (http://www.radiobras.gov.br) is a Brazilian public company that aims to establish a communication channel between the departments of Federal Government and the Brazilian society. By this way, Radiobrás pursues to universalize the information regarding the acts and facts of the Federal Republic of Brazil. Its objectives are:

(a) To publish the accomplishments of the Federal Government in the economy, social politics and to spread out abroad adequate knowledge of the Brazilian reality, as well as implanting and operating senders and exploring services of broadcasting of the Federal Government;

(b) To implant and to operate its repeating networks and retransmission of broadcasting, exploring its services, as well as promoting and stimulating the formation and the training of specialized staff necessary to its activities;

(c) To gather, elaborate, transmit, and distribute, directly or in cooperation with other social communication entities, news, photographs, bulletins and programs concerned to acts and facts from the Government and other issues of political, financial, civic, social, sportive, cultural and artistic nature, by means of graphical, photographic, cinematographic, electronic or any other vehicle;

(d) To distribute the legal publicity from the entities related directly or indirectly to the Government;

(e) To perform other activities assigned to it by the Chief Ministry of State of the Government Communication Secretariat of the President of the Republic.

Paper from: *Data Mining III*, A Zanasi, CA Brebbia, NFF Ebecken & P Melli (Editors).
ISBN 1-85312-925-9

By means of this project, Radiobrás aims to obtain indicators related to the distribution of news by subject, the diffusion of news abroad, to estimate the distribution of news along the government departments, to check the news contents regarding to the Communication Secretariat. For this purpose, efforts will be focused in measuring the amount of news by kind, period and main topics, in extracting the concepts produced and propagated by the agency based on clustering analysis, determining the amount of news about acts and facts of the Government, and studying the degree in which Radiobrás is achieving its objectives. This work meets these objectives by *(a)* determining the most important words in the issued news, *(b)* determining the main correlation among the news and the keywords that compose each cluster, *(c)* separating news by groups, *(d)* pointing out the most representative words, *(e)* discovering the main concepts from the clustering analysis, and *(f)* elaborating statistics about the news by time, subject and quantity.

3.2 Data understanding

The data were obtained from the public repository of the agency. Each text file corresponds to specific news. Corrupted, control files and news files in foreign language were discarded.

We considered just the news produced in 2001. Figure 3 shows the monthly production of news in this year.

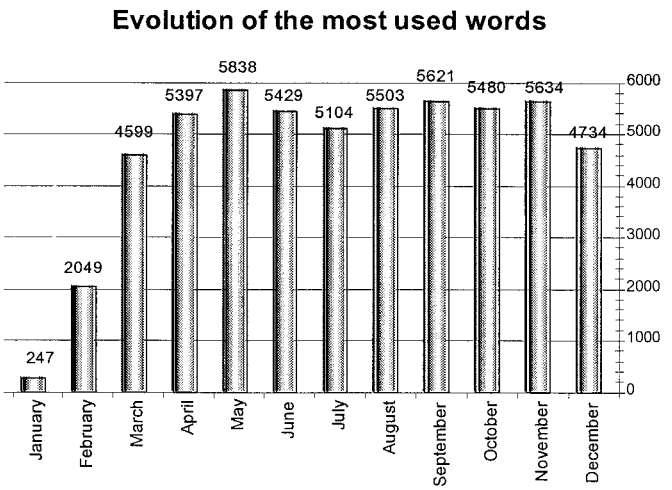

Figure 3 – Monthly production in 2001

3.3 Data preparation

The selected news, that includes releases, complete notices, guidelines, presidential agenda, events in course and photos are prepared according to the cycle depicted in Figure 4.

In this phase we found that the production from January and February should consider as outliers and consequently, discarded. This happened due to problems of importing texts from the repository that caused loss of records.

Paper from: *Data Mining III*, A Zanasi, CA Brebbia, NFF Ebecken & P Melli (Editors).
ISBN 1-85312-925-9

3.4 Data modeling

We carried out the work in this phase by describing and summarizing the data, and, then, segmenting the set of texts. It was applied Tan [5] approach, in which two steps are performed: *(a)* text refinement, which corresponds to transforming the text from free form to an intermediary form, and *(b)* knowledge extraction, corresponding to the data mining itself. An example of data description and summarization is shown in Figure 5. Results from segmentation can be seen in Figures 6 and 7. To induce the clusters we applied the Eurekha [6] tool, choosing the Star option. The overall process of clustering, required for segmentation, is depicted in Figure 8.

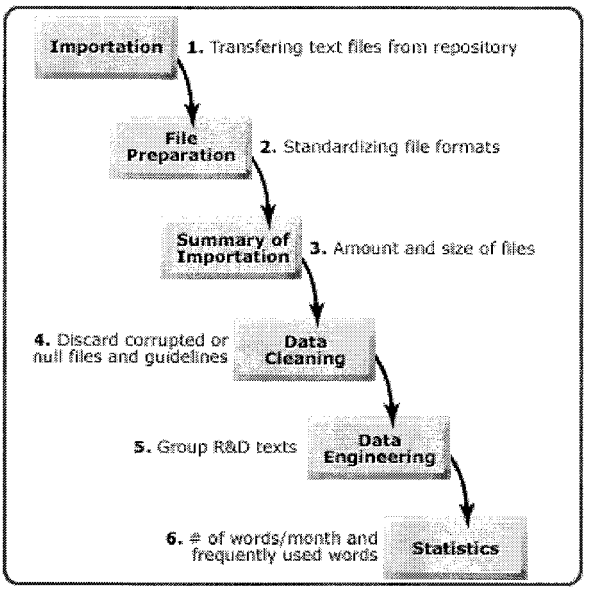

Figure 4 – Data preparation

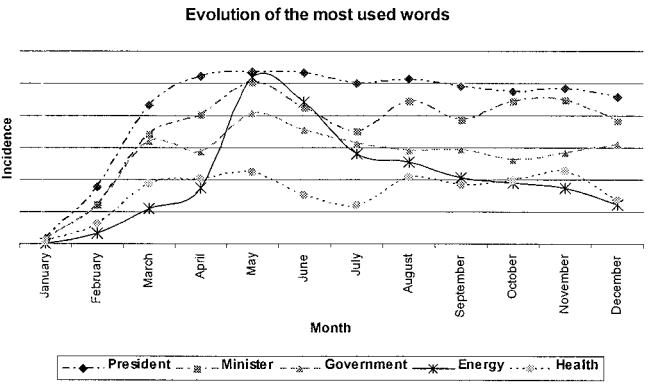

Figure 5 -- Incidence of most used words (2001)

Paper from: *Data Mining III*, A Zanasi, CA Brebbia, NFF Ebecken & P Melli (Editors).
ISBN 1-85312-925-9

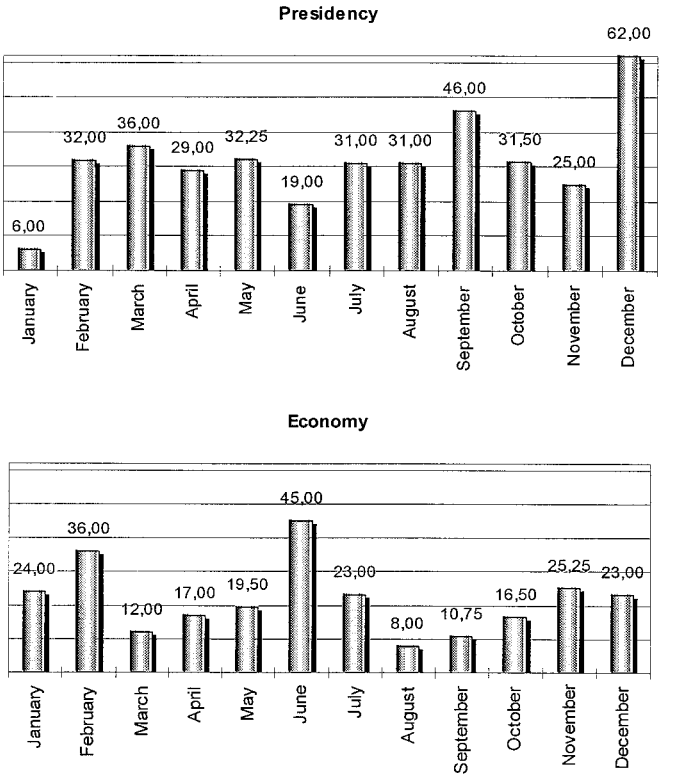

Figure 6 – Examples of categories: Presidency and economy

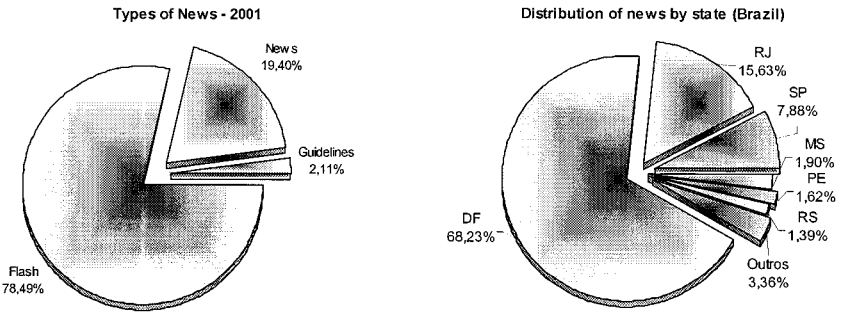

Figure 7 – Kind of news and geographic distribution

Paper from: *Data Mining III*, A Zanasi, CA Brebbia, NFF Ebecken & P Melli (Editors).
ISBN 1-85312-925-9

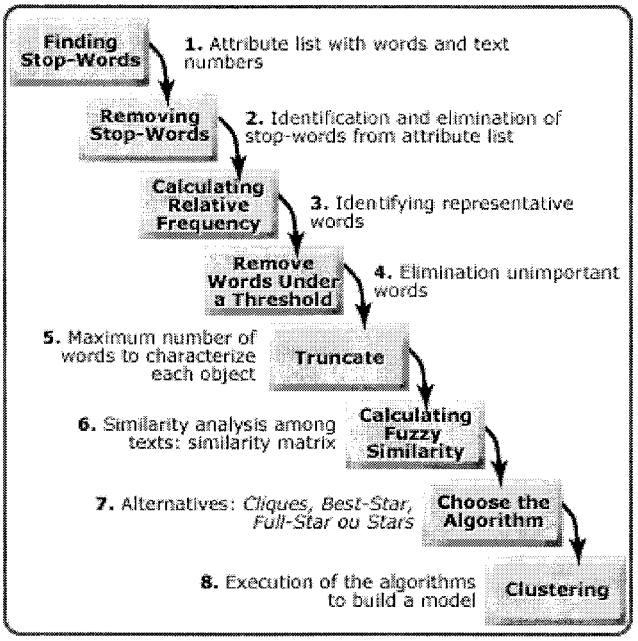

Figure 8 – Methodology for clustering in KDT

3.5 Model evaluation

The clusters found in the previous phase were analyzed by an expert in order to discover some meaning in them. After that, a categorization by subject was carried out. This categorization was perfomed by applying the methodology introduced by Halliman **[3]** that combines cluster analysis with background knowledge.

3.5.1 Results

After analyzing the categories found, five major areas were identified: *(a) Presidency of Republic* - 30% (president, Fernando, Henrique, Cardoso), *(b) Economy* - 21% (central bank, monetary values, inflation, stock exchange, dollar rates, interests, national treasure), *(c) Meteorology* - 21% (time forecast, cloudy, partially, rain), *(d) Development* - 8% (energy, monetary values, state companies, investment and development) and *(e) Politics* - 7% (parties acronyms, house of representatives, senators, ministry names, senators names). It is important to emphasize that this classification was just a feeling before the present analysis and now have a sound rationale. Almost all news does not have any label that could be used as category.

Other less frequent categories, that total 13% of the whole text set, were also listed. Next this categories as well as their corresponding key words, are described: *Education* (school census, university, national school evaluation), *Health* (AIDS, HIV, clone, generic medicines, hospital, cholesterol, medicine),

Paper from: *Data Mining III*, A Zanasi, CA Brebbia, NFF Ebecken & P Melli (Editors).
ISBN 1-85312-925-9

International (UN, WWF, El Salvador, Mercosul, Palestine, Israel, New York, attempted against, towers, United States), *Security* (federal police, antidrugs), *Providence* (INSS, social security, deadline), *R&D* (Genoma, technology), *Justice* (court, federal, justice, trial, Indian, Galdino), *Environment* (environment, birds, ISO, Amazon), *Agriculture* (INCRA, agrarian reform, IBAMA, soil), *Culture* (carnival, art, museum, exhibition, orchestra, symphonic, beautiful, winter festival), *Transportation* (airports, police, bus station, subway, conference, traffic), *Sports* (INTECOM, ECT, soccer, Nike, CBF, sets, marathon, Vasco, Gama, Cruzeiro, Goiás, olympic games), *National* (quality, price, meat, accident, P-36), *Work* (woman, agreement, rural, work, infant, forum, SENAC, SEBRAE).

For a better understanding of the categories, it was created many graphs like in Figure 5 that shows the most used words during 2001. They were interpreted by an expert that issued the interpretations below.

Evaluation of Figure 5. The constant use of the words "president", "government", and "minister" suggest the approach of actions taken by central administration. The frequency of the word "healthy" increases as the government and the Healthy Department make public vaccination and drug prevent campaigns (e.g., against AIDS). The most scored word was "energy" pointing out the effort employed by the central administration to deal with the lack of energy in the country in a certain period.

Evaluation of Figure 6. The categories "presidency", "politics", "development", and "economy" meets the agency objectives regarding to the coverage of acts and facts generated by the central administration. They also reflect that, in each month, there are coincidence between the news and important facts. We can mention, for example, financial crisis and "economy", development and "blackout campaign", war and terrorism in USA and "International" and "Security", strike in the metro, bus and trains and "transportation", educational campaigns and "education".

Evaluation of Figure 7. Almost all news are flashes (IT-Internet) that do not bring any other information but the pure text (e.g., it is not informed the news focus). It does not allow a more precise evaluation regarding to the distribution inside the slices. It is also possible to verify that the biggest amount of news come from Distrito Federal (DF), Rio de Janeiro (RJ) and São Paulo (SP). It agrees with the fact that Brasília and Rio de Janeiro are the headquarters of many public departments. In this graphic we can see the presence of Pernambuco, mainly due to the news related to the so-called "polygon of marijuana".

3.5.2 Evaluation of results

Considering the success criteria defined by the administration, the results were considered to fulfill the organization's expectations. Actually, since the application raised new questions, the user decided to keep the studies in order to these new questions.

Paper from: *Data Mining III*, A Zanasi, CA Brebbia, NFF Ebecken & P Melli (Editors).
ISBN 1-85312-925-9

3.6 Development

The application has shown to be an important alternative to develop an institutional self knowledge useful for a better management both internally and externally.

The results available for this purpose include: *(a)* main subjects approached in the news, *(b)* monthly production, *(c)* geographical distribution of news' sources, *(d)* clusters and their most important words, *(e)* different kinds of subject categorization, and *(f)* comparison between issued news and current national and international facts.

By knowing the subjects approached in the issued news, the heads of Radiobrás has developed an effective view of the role it is playing in the society, being able to correct any deviation in accomplishing its mission.

4 Conclusion

We departed from the fact that, although 80% of a company's information is contained in text documents, only 7% of KDD applications are developed to process pure or marked text. A fair belief is that this huge amount of information hides useful knowledge that could lever the organization to a better position in the market. With these facts in mind we decided to investigate the low interest in text mining enumerating, initially, some hypotheses related to usability, few reported experiences, and the lack of methodology. In this work we focused our attention to the methodological hypothesis, carrying out our research by means of a case study in a Brazilian news agency. We applied the CRISP-DM methodology, which was originally elaborated to drive data mining applications (that process structured data). The methodology was completely adequate to develop text mining application, as the obtained results can show.

It became evident that the methodological hypothesis should be disregarded, opening a research opportunity to study other hypotheses.

References

[1] AGRAWAL, R. "Data Mining: Crossing the Chasm", Invited talk at the 5th ACM SIGKDD Int'l Conference on Knowledge Discovery and Data Mining (KDD-99), San Diego, California, August 1999. http://www.almaden.ibm.com/cs/quest/PUBS.html (15/05/2001)

[2] CHAPMAN, P., KERBER R., CLINTON J., KHABAZA T., REINARTZ T., WIRTH R. – "The CRISP-DM Process Model", Discussion Paper, 2000. http://www.crisp-dm.org (08/07/2001)

[3] HALLIMAN, C. "Business intelligence using smart techniques: environmental scanning using text mining and competitor analysis using scenarios and manual simulation", Information Uncover, Houston, 2001.

[4] NUGGETS® "KDnuggets.com (KD stands for Knowledge Discovery) is the leading source of information on Data Mining, Web Mining, Knowledge Discovery, and Decision Support Topics". http://www.kdnuggets.com/polls/data_mining_techniques.htm (21/08/2001)

Paper from: *Data Mining III*, A Zanasi, CA Brebbia, NFF Ebecken & P Melli (Editors).
ISBN 1-85312-925-9

[5] TAN, A.-H. "Text mining: The state of the art and the challenges", Kent Ridge Digital Labs, 1999. http://textmining.krdl.org.sg (23/08/2001)
[6] WIVES, L. K. "Um Estudo sobre Agrupamento de Documentos Textuais em Processamento de Informações não Estruturadas Usando Técnicas de Clustering" MSc Dissertation, Porto Alegre (Brazil), PPGC/UFRGS, 1999.