Concept-Based Knowledge Discovery in Texts Extracted from the Web

Stanley Loh
Univ. Católica de Pelotas (UCPEL)
Univ. Luterana do Brasil (ULBRA)
Univ. Federal do Rio Grande do Sul (UFRGS)
sloh@zaz.com.br

Leandro Krug Wives
Univ. Federal do Rio Grande do Sul (UFRGS)
wives@inf.ufrgs.br

José Palazzo M. de Oliveira
Univ. Federal do Rio Grande do Sul (UFRGS)
palazzo@inf.ufrgs.br

Address
Programa de Pós-Graduação em Computação
Instituto de Informática
Universidade Federal do Rio Grande do Sul
Avenida Bento Gonçalves, 9500
Bloco IV, Prédio 43412 - Campus do Vale
Porto Alegre - RS - 91501-970
BRASIL

ABSTRACT

This paper presents an approach for knowledge discovery in texts extracted from the Web. Instead of analyzing words or attribute values, the approach is based on concepts, which are extracted from texts to be used as characteristics in the mining process. Statistical techniques are applied on concepts in order to find interesting patterns in concept distributions or associations. In this way, users can perform discovery in a high level, since concepts describe real world events, objects, thoughts, etc. For identifying concepts in texts, a categorization algorithm is used associated to a previous classification task for concept definitions. Two experiments are presented: one for political analysis and other for competitive intelligence. At the end, the approach is discussed, examining its problems and advantages in the Web context.

Keywords

Knowledge discovery, data mining, information extraction, categorization, text mining.

1. INTRODUCTION

The Web is a large and growing collection of texts. This amount of text is becoming a valuable resource of information and knowledge. As Garofalakis and partners comment, "*the majority of human information will be available on the Web in ten years*" [21]. To find useful information in this source is not an easy and fast task. People, however, want to extract useful information from these texts quickly and with low cost.

The heterogeneity and the amount of sources may lead to the information overload problem, which happens when we have too much information available that we cannot manage. To minimize the overload and to help people to extract information from texts has emerged the novel area called *Knowledge Discovery in Texts* (KDT) [15], which concerns the application of *Knowledge Discovery in Databases* (KDD) techniques over texts. KDD is the "*nontrivial extraction of implicit, previously unknown, and potentially useful information from given data*" [19]. However, the most researches in KDD work on structured data (like in a database) and cannot be applied over textual data directly.

Feldman and partners [15] [16] [17] face the problem of applying KDD tools over keywords that are assigned to texts as attributes. These mining techniques use statistical analysis to discover association rules and interesting patterns over keyword distributions and associations. To perform the KDT process, keywords should be previously assigned to texts. Authors do not discuss the way in which keywords are assigned to texts, suggesting that this assignment may be done manually by humans or automatically by software tools.

By other side, Lin and partners [31] use terms automatically extracted from texts to categorize documents and to find associations. The most frequent terms are assigned as keywords (attributes). However, when analyzing terms, problems arise due to the *vocabulary problem*, discussed in [5], [7] and [20]. The language use may cause semantic mistakes due to synonyms (different words for the same meaning), polysemy (the same word with many meanings), lemmas (words with the same radical, like the verb "to marry" and the noun "marriage") and quasi-synonyms (words related to the same subject, object or event, like "bomb" and "terrorist attack"). For example, a murder may be described with terms like "murder" or "homicide". If analyzing only the terms (assigned to or extracted from texts), the discovery process may be misled by semantic gaps.

Another interesting approach is to apply KDD techniques after the use of Information Extraction (IE) techniques, which transform information present in texts into a structured database [10]. When textual information is structured into a database, we can do useful analyses only possible with Database Management Systems [33]. For example, using associative techniques, one can discover relations between items examining transactions in a database. In [21], we can see an approach that uses associative techniques over Web pages structures. In this case, URLs are extracted from pages to represent items in a database. However, Etzioni [14] advises: "*HTML annotations structure the display of Web pages, but provide little insight into their content*". Besides that, texts may even be published without titles, keywords, links or author information, making HTML tags useless.

In some cases, IE has shown to be feasible in exploring textual content. Soderland [43] extracts information about weather forecasting in Web texts. Etzioni [14] cites some successful cases of IE applications using "wrappers" (Web information extractors). Although the promising results in IE, unfortunately the "*majority of today's IE systems rely on hand-coded wrappers to access a fixed set of Web resources*" [21]. This means IE systems are very domain dependent, being useful only for specific applications [11] or working only with a special class of document types. Besides that, to create such systems, a lot of knowledge about the domain is necessary (knowledge engineering), examining text styles and how information is encoded into natural language phrases [8]. Mattox and partners [33] conclude that semantic knowledge about the domain (like ontologies) is essential to IE and that there is a non-trivial effort to generate wrappers, even with tools. When the access to information is infrequently, it is not worth the effort.

This paper presents an approach for performing knowledge discovery in texts extracted from the Web, through the analysis of high level characteristics, minimizing the *vocabulary problem* and the effort necessary to extract useful information. Instead of applying mining techniques on attribute values, terms or keywords labeling texts, the discovery process works over concepts extracted from texts. Concepts represent real world attributes (events, objects, feelings, actions, etc) and, as seen in discourse analysis, they help to understand ideas and ideologies present in texts. The approach combines an automatic categorization task with a mining task. Categorization task identifies concepts present inside texts, without needing too much labor. Mining task discovers patterns by analyzing and relating concept distributions in a collection. A previous classification task is needed to create concept definitions.

For a better communication, we distinguish classification from categorization. Classification is the process of inducing a model or description for each class in terms of its attributes [21]. This model is then used to identify the class of future elements. Categorization refers to the identification of categories, themes, subjects or concepts present in texts. Lewis and Hayes [28] give a different definition for text categorization: "*the classification of units of natural language text into predefined categories*", and Wiener and partners [46] call this problem as "*topic spotting*". Some authors tend to view categorization as part of the classification. Thus, in some papers classification and categorization are used as synonyms. Here, we use these terms as different meanings.

The main goals of the proposed approach are: (a) to do discovery upon concepts instead of words or attribute values, allowing the user to find ideas, ideologies, trends and intentions present in texts; (b) to minimize the effort necessary to identify concepts in texts and to perform knowledge discovery; (c) to find interesting patterns in textual collections using simple statistical techniques; (d) to allow users to perform *ad hoc* discovery (with ill-defined goals) without having to expend time and effort creating formal models.

Applications of the concept-based KDT approach include (but are not limited to): discourse analysis (looking for intentions and structures in textual expressions), sociology (themes and ideas present in a textual study), analogy (same concepts present in different discourses), health research (searching for relations between symptoms present in textual records) and competitive intelligence (strategies used by different companies).

The section 2 presents a general overview of the approach. Subsections discuss each task of the concept-based KDT process. In section 3, results from two experiments are presented. Section 4 discusses the experiments and the final section evaluates the approach.

2. THE CONCEPT-BASED APPROACH FOR KDT

Although many researchers use the term "concept", it is difficult to see a formal definition of what "*concept*" is. Looking for a definition in dictionaries, we find that "concept" is an "idea, opinion, thought". This confirms the general and intuitive idea that concepts are used to explore and examine the contents of talks, texts, documents, books, messages, etc. Chen and partners [6], for example, use concepts to identify the content of comments in a brainstorming discussion. In Information Retrieval, concepts are used with success to index and retrieve documents. Lin and Chen [30] comment "*the concept-based retrieval capability has been considered by many researchers and practitioners to be an effective complement to the prevailing keyword search or user browsing*". In this case, its main advantage is to minimize the vocabulary problem.

According to [44], concepts belong to the extra-linguistic knowledge about the world. Sowa [44] states: "*the concepts expressed by a language are determined by the environment, activities, and culture of the people who speak the language*". So, the use of concepts depends on who is doing that, for what purpose and in what context. For example, intending to analyze discourses of politicians, one may want to identify concepts like "progress", "problems", "investments", "money", "corruption", etc. By other side, in a psychiatric environment, concepts may be "violence", "drugs", "suicide", "death", etc. Soderland [43] defines concepts inside the weather forecasting domain. Each weather condition ("cloudy", "fair", "precipitation") is a concept and has its own definition. Also there are concepts to time and days.

The way in which concepts are represented in formalisms is suited to particular viewpoints. There are many and different approaches to express mental models. However, we are interested in a simple structure that allows us to represent real world objects, events, thoughts, opinions and ideas, easily and with a certain degree of quality for the discovery process. Following [4], [6] and [42], we use the *vector space model* to represent concepts internally. So each concept is stored as a set or vector of terms. Although it is possible to represent concepts as a network [6] or as an ordered list of terms [9], we have decided to use a non-ordered vector without links, assuming that all terms inside a concept description are related to each other in a same degree. The decision for this structure is to simplify classification and categorization tasks.

Terms in a concept (its descriptors) may include synonyms, quasi-synonyms, lexical variations, plural, verb derivations, semantic related words, etc. Associated to each term in a vector there must be a weight, ranging from 0 to 1 and describing the relative importance of the term to indicate whether a concept is present in a text. According to [4], this approach is better than the binary model since term count information leads to higher accuracy. Terms work like tokens, so it is not necessary that a term have a universal meaning, being possible to use proper nouns and abbreviations, even if they are meaningful only for the domain people (from here, we use "*term*" and "*word*" like synonyms).

Each concept has only one set as a descriptor, but one term may be present in more than one descriptor set. Currently, only single words are allowed. Although we know pairs of terms can give better results [1], our choice for using single words is due to computational limitations. Using simple representations of concepts reduces the time to perform classification and categorization. We expect to achieve good performance by the context analysis. The *fuzzy* function used in the categorization task tends to reward the presence of more than one word. Besides that, according to [1], using only pairs bring poor results, while single words alone are relatively successful.

One assumption is that the concept-based approach tends to minimize the *vocabulary problem* because concepts may be expressed with different words, as in a semantic expansion approach (see benefits of the semantic expansion technique for Information Retrieval in [3], [24] and [45]). So the efficiency of identifying concepts within a text is higher because more terms are covered. Chen [5] argues that people tend to use different terms to describe a similar concept. Furnas and partners [20] discuss the effectiveness of an "*unlimited aliasing*" strategy, which allows unlimited number of aliases for objects, to minimize the *vocabulary problem*. When examining Information Retrieval strategies, Bates [2] found that "*for a successful match, the searcher must somehow generate as much 'variety' in the search as is produced in indexing"*. This kind of *redundancy* allows identifying overlapping words similarly to how people express concepts and ideas (text authors and someone performing knowledge discovery in texts).

Another assumption is that the effort for concept definition and identification can be reduced. Thus knowledge engineers are not necessary and users do not need to expend too much effort and time to define models and rules to extract information, as when using *thesauri*, ontologies and natural language processing. Feldman and Dagan [15] defend the use of simple structures because they allow tasks to be performed with computer aids and with low costs.

In summary, we may compare the KDT approach against the KDD phases suggested in [22]:

a) understanding the application domain and the goals of the data mining process: user must define which concepts are interesting (first part of the classification task);

b) acquiring or selecting a target data set: texts must be gathered, using IR tools or in a manual way;

c) integrating and checking the data set: in our approach, texts must be saved in individual textual files (*.txt), no other validation is done;

d) data cleaning, preprocessing and transformation: concepts must be described (second part of the classification task) and texts need to be analyzed and stored in the internal format (vectors of words with weights representing the relative frequency), after eliminating *stopwords*, following suggestion of [4];

e) model development and hypothesis building: identifying concepts in the collection (the categorization task);

f) choosing suitable data mining algorithms: the application of the statistical techniques (mining task);

g) result interpretation and visualization: humans must interpret the findings;

h) result testing and verification: redoing the process or some stages to validate the discovered knowledge;

i) using and maintaining the discovered knowledge: done by humans.

The main tasks of the approach (categorization, classification and mining) are discussed in the next sections.

2.1 How to Identify Concepts Inside Texts (Categorization)

According to [15], one kind of information extraction is the categorization of a text by meaningful concepts. The goal of the categorization is to identify concepts present in texts. However, documents do not have concepts explicitly, but rather words [1]. Once concepts are expressed by languages (words and grammars) [44], it is possible to identify them in texts by analyzing phrases.

Instead of using complex Natural Language Processing (NLP) to analyze syntax and semantics, our approach is based on a simple technique. We believe concepts may be identified by cues (terms). Using a *fuzzy* reasoning about the cues found in a text, we can calculate the likelihood of a concept being present in that text. This is a kind of Information Extraction, except that it is not necessary to fill fields with values. The extraction only needs to identify the presence of concepts. We consider that this approach is under the statistical NLP paradigm according to the definition in [25], since it uses frequency counting and probability theory. However, syntax analyses are not done.

The categorization algorithm follows Rocchio's one [41], since it uses a prototype-like vector (a centroid) to represent each class/category (in our case, the concepts) and evaluates the membership of an element (the text) in a class using a similarity function that calculates the distance between the element and each centroid. The choice for this algorithm is due to its simplicity, ease to implement and relative efficiency, according to [9]. Ragas and Koster [39] carried out experiments using four different algorithms and found that Rocchio's and Bayes algorithms achieved better results. They suggest a combination of both. The main disadvantage of these algorithms is that the context of words (near words) does not influence the categorization [9]. This may cause problems since the context may change the meaning of a word or the interpretation of a phrase (for example, "*is*" and "*is not*").

The algorithm starts comparing all texts against each concept, assuming that concepts were defined early and texts previously represented in the internal format. The comparison is done through a *fuzzy* reasoning process, following [49] and [37]. Weights of common terms (those present in both text and concept) are multiplied. The overall sum of these products, limited to 1, is the degree of relation between the text and the concept, meaning the relative probability of the concept presence in the text or that the text holds the concept with a specific degree of importance. Terms that are not present in the intersection of the representations are not counted because concept descriptors may be using synonyms.

The fundament behind this process is that each word of a concept contributes with certain strength to the presence of that concept. Strong indicators may receive higher weights in the concept definition (as will be discussed in the next subsection). Indeed, we are working with signs under uncertainty. This is like the relevancy index proposed in [40] whose definition is "*a collection of features that, together, reliably predict a relevant event description*". Similarly, Morris [36] distinguishes between indicator signs and characterizing signs. The first ones point to a specific object or element, while the last ones restrict elements in a set. Alike this thought, McCarthy [34] comments the use of approximate concepts. According to this author, there are sufficient and necessary conditions to certify the presence of a concept. Sufficient conditions (SC) works like the implication SC → CONCEPT, while necessary conditions (NC) are like CONCEPT → NC.

In our approach we consider that terms are characterizing signs and necessary conditions. Terms indicate the presence of a concept with a degree of certainty (TERM → CONCEPT). So the *fuzzy* reasoning must evaluate the likelihood of a concept to be present in a text, analyzing the strength of its indications. The process is like an abductive reasoning. According to [23], in a deduction, if "A → B" and "A is truth" then we can infer "B is truth". In abduction, if "A → B" and "B is truth" then "A is a probable cause for B being truth". That means if words that describe a concept appear in a text, there is a high probability of that concept being present in that text. The decision concerning if a concept is present or not depends then on the threshold used to cut off undesirable degrees. Riloff and Lehnert [40] evaluated three methods for identifying concepts in texts. Two of them consider that a concept is present if and only if there is a keyword or key phrase in the text, but they are prone to false hits due to the *vocabulary problem*. The third method analyzes the context, using a relevancy degree. They concluded that the choice for one of these methods depends on the collection and language characteristics.

Our approach uses the threshold to decide whether a concept is present or not. As the user may set this threshold, it is possible that only one term indicates the concept presence. However, the more indicators are present, the more likely the concept is present. The decision is then done by the context analysis. Chakrabarti [4] believes that "*using a statistical method for text implies that the learned rules will not be dependent on the presence or absence of specific keywords*". This threshold may be chosen in a training session, before the categorization task.

2.2 How to Acquire Concept Definitions (Classification)

The classification task is responsible by generating concept definitions, that is, the choice of concepts and the description of each concept (terms and their associated degrees of relevance).

Chen and partners [7] suggest to use either an existing controlled vocabulary (like dictionaries, *thesauri* or ontologies) or to automatically generate one. The main problem with *thesauri* is that they are usually very domain dependent and in some cases do not support slight variations because they do not have sufficient vocabulary coverage for all potential applications or specific user groups. Yang and Chute [47] reported problems with a medical *thesaurus*, because physicians used other words in their daily practice. In its turn, ontologies like WordNet [35] fail to include proper nouns. Although Liddy and partners [29] have demonstrated the benefits from using dictionaries, sometimes they do not include important semantic relations. In a previous study (not published yet), we found that definitions present in Webster-like dictionaries use too many general words, as for example in "*soccer = ball game played with feet, disputed by two teams with eleven players each...*" Examining the presence of the *soccer* concept in newspapers, we found that the listed words are not so frequent. By the other side, preexisting vocabularies may not be appropriated to the user's needs (lack of specificity or missing interesting concepts).

The automatic generation of a controlled vocabulary is a learning process [4] and can be done through either a supervised or an unsupervised process. The problem of the supervised process is that a high-quality sample of data must be available [1]. To find such a sample in an environment like the Web is a difficult task and demands a lot of time and effort.

For an unsupervised learning, Etzioni [14] suggests the clustering technique, which does not require labeled inputs. Fisher [18] states that the clustering process "*accepts object descriptions (events, observations, facts) and produces a classification scheme over the observations*". According to the same author, "*a learning of this kind is referred to as learning from observation (as opposed to learning from examples)*". However, classes are identified and created apart from the user's interest and this may not be appropriate to the application goal.

As we need a method that could be efficient (not necessarily the best) but mainly having low cost in terms of time and effort, we have chosen a manual process helped by dictionaries and software tools. We believe that, in the Web environment, users have *ad hoc* needs and do not want to spend time in computations or defining formal models. Besides that, in a manual process, concepts may be pruned to the users' interest.

However, our suggestion is that preexisting vocabularies, such as a *thesaurus* or a technical dictionary, if available, should be used to minimize the effort in this task. Besides that, a general dictionary (like a Webster's) may help the user to find synonyms. Bates [2], for example, proposes the use of a domain-specific dictionary to expand the user's vocabulary. Yang and Chute [47] showed the efficiency of the combination of a technical dictionary, like the CID (International Code/Classification of Diseases) in the medical area, augmented with synonyms specific of the domain.

Also it is important to examine some examples of the language style. In this way, software tools can play an important role, helping users to identify words used in the collection. According to [25], little samples can bring good results in some particular cases. Besides that, software tools minimize the knowledge acquisition bottleneck (according

to [7], the cognitive demand required of humans to create controlled vocabularies).

As each word in a concept must have an associated value of importance, the user must define them too. Since it is difficult to assign numeric values, *fuzzy* linguistic variables may be used [49]. However, we use a software tool to help in the weight definition. This tool shows all the words present in a set of texts and the frequency of each one. Thus, user can examine a sample of the collection and verify which words are more common and in which context they occur. Lagus and Kaski [26] state that a good descriptor must characterize some outstanding property and Salton and McGill [42] suggests that good descriptors are those that are frequent inside a text but infrequent in the whole collection (small inverse frequency). So, we suggest to assign small values to generic words or those present in more than one concept (or even eliminating this kind of word) and to assign higher values (for example, 1) to those that appear only in one concept description.

2.3 Using Statistical Techniques on Concepts (Mining Task)

The approach analyzes concept distributions to discover interesting patterns. This is like the IE+KDD paradigm, where IE is performed early to extract text attributes and KDD techniques are then used over these attributes. The difference is that we perform analyses over concepts instead of words or values (concepts work as text attributes). The approach may be considered under the probabilistic and statistical paradigm according to [32], since it is based on the distribution of variables in the collection. Following we discuss the techniques used for the mining task, assuming classification and categorization are finished. These techniques do not consider the degree of relation between a text and a concept (how much a concept is present in a text). We assume that it is only important to know if a concept is present or not inside a text.

The first technique used is the key-concept listing, which analyzes concept distributions over the collection. We have a software tool that extracts a concept-based centroid of a collection. After the categorization task, each text has associated to it a list of concepts with relative degrees. For each concept, the tool counts the number of texts to which the concept is assigned. The degrees are not considered in this step, because it does not matter how much a concept is present in a text but only if it is present (regarding that categorization must have cut off undesirable degrees according to the chosen threshold). This technique allows for finding which dominant themes exist in a collection or in a single text. Also we can compare one centroid to another, to find common themes or changes between sub-collections. Another possible usage is to find differences between sub-collections or concepts present in only one text. We followed Feldman and Dagan's [17] suggestion for examining distributions that differ significantly from the full collection, from other related collections or from collections in a different time.

The second technique is the association or correlation. It discovers associations between concepts and expresses these findings as rules in the format X ➔ Y (X may be a set of concepts or a unique one, and Y is a unique concept). The rule means "*if* **X** *is present in a text, then* **Y** *is present with a certain confidence and a certain support*". Following the definitions of [31] and [21], *confidence* is the proportion of texts that have X AND Y in relation to the number of texts that have only X, and *support* is the proportion of texts that have X AND Y in relation to all texts in the collection. Rules allow predicting the presence of a concept according to the presence of another one. Complex rules may be discovered with human intervention. So the precedent part of a rule may be a combination of concepts and/or words, such as WORD_1 AND WORD_2 AND CONCEPT_1 AND CONCEPT_2 => CONCEPT_3. This kind of rule is found using intermediary retrieval tasks, to select sub-collections where some words are present. In the next section, some examples will be explained.

3. EXPERIMENTS

We carried out some experiments with textual collections extracted from the Web to validate the concept-based approach presented here. In this paper, we discuss two experiments, one in a political analysis context and another for competitive intelligence (business intelligence) analysis over Text Mining tools.

The way in which the concepts were defined (classification task) was different in each experiment. Under a certain viewpoint, we can say that these two styles are complementary. In the political experiment, concepts were defined through an exhaustive analysis of words used in the collection. That means that we examined every word present in more than one text and classified it into a concept, resulting in a set of 104 concepts. Each word being examined could be classified into an existing concept or generate a new one. *Stopwords* and general terms were eliminated before this examination. By another side, in the competitive experiment, interesting concepts (according to the experiment goal) were first selected and then defined and refined through the examination of words present in the collection, giving a total of 24 concepts.

These experiments show how concepts may be defined. The first alternative generates a bigger set but helps the user to find which concepts are present in the collection. Besides that, this kind of definition intends to cover all relevant words in the collection. The second alternative, by other side, narrows the set of concepts to only those relevant to the specific goal. This allows the user to choose previously the concepts of his/her interest and therefore tends to generate a smaller set. Classification task took less than 30 minutes in the first experiment and about 10 minutes in the second experiment. Both tasks were supported by a software tool that analyzed the words present in texts using a Pentium II 400 MHz with 64 Mbytes of RAM.

In these experiments, we used as interestingness measures a confidence threshold of 80% and a minimum support equals to 60% or 5 texts. Feldman and Hirsh [16] suggest a minimum support of 5 documents and a confidence threshold of 10% for the association rules.

3.1 Political Experiment

The goal of this experiment was to extract knowledge about what press is or was telling about the mayor of a big city in Brazil. To represent the press, an online newspaper was used. Texts were written in Portuguese. Using a local search engine and the mayor's name, we gathered 180 texts published in 1997 and 178 texts published in 1999, forming two sub-collections for a later comparison.

Examples of concepts definitions are: *"crimes"* = {crime, crimes, fraud, fraudulent, illegal…} and *"elections"* = {election, elections, term, reelection, voter, elected, electorate,…}.

The most interesting patterns and their interpretation are listed below according to the mining technique used. It is important to say that the mayor is under investigation by the Department of Justice on charges of corruption since the beginning of 2000.

Associative rules (association technique)

a) *drug traffic* ➔ *politicians* (confidence = 93.3%, support = 14 documents)

This means that when a press release dealt with "drug traffic", the name of a politician was cited too. We do not know why the names were cited, whether in a favorable way or not (accusing or accused), unless we read the texts. But this pattern allows us to conclude that the drug problem achieved the political sphere and a high importance degree, when the mayor is involved.

b) *loans* ➔ *politicians* (confidence = 82.1%, support = 23 documents)

This pattern was discovered in the 1997's sub-collection and means that references to "loans" of any kind involved the name of a politician. As the previous finding, it does not allow us to conclude the cause, but we may infer politicians are involved asking for, releasing, criticizing or receiving loans.

c) an interesting combination of 2 patterns,

(1) *loans* ➔ *politicians* (confidence = 82.1%, support = 23 documents)

(2) *education* ➔ *politicians* (confidence = 64.2%, support = 27 documents)

raised the hypothesis of a connection between "education" and "loans". When examining the direct relation between the two concepts, we found

(3) *education* ➔ *loans* (confidence = 4.7%, support = 2 docs)

(4) *loans* ➔ *education* (confidence = 7.1%, support = 2 docs).

These results lead us to conclude that there is not a direct relation between "loans" and "education". Probably, rules (1) and (2) happen in different subsets. However, when analyzing these two concepts together, the following rule was discovered,

(5) *loans* AND *education* ➔ *politicians* (confidence=83,3%, support=5 docs),

allowing for the conclusion that "politicians" are involved when "loans" and "education" are cited together, perhaps influencing decisions. However, not all "loans" with politicians' involvement are related to "education", because only 17,2% of the cases involving "loans" and "politicians" have "education", conform the next rule

(6) *loans* AND *politicians* ➔ education (confidence=17,2%).

Concept distributions (key-concept listing technique)

Analyzing the whole collection (358 texts), we found as the main themes: *politicians* (140 texts, 39.1%), *crimes* (117 texts, 32.6%) and *elections* (105 texts, 29.3%). From that, it is possible to infer that references to "crimes" are common when the mayor is cited, since the theme appears with frequency similar to political themes.

Comparing the distributions of concepts in 1997's to 1999's, some interesting observations arose:

a) comparing the two sub-collections, it can be observed that the 1997's sub-collection had a dominant focus (the presence of politicians associated with the mayor), while in 1999 the themes had a balanced distribution;

b) the weight of the "*elections*" concept rose from 25% (1997) to 33.7% (1999), possibly due to the nearing election in 2000 (the mayor's term finishes in December 2000);

c) the "*debts*" concept reduced its participation from 1997 to 1999, meaning there was a reduction in "debts" or the press changed its interest to other topics;

d) a particular interest exists to observe the distribution of names of people: an interesting pattern is that the presence of the mayor's main associate went down from 1997 to 1999; using our knowledge about the domain, we interpret this event as consequence of the political separation between them, during this time gap; also we can observe over this special sub-collection (when both politicians are cited together) that the main focus has changed from "debts" (40.7% => 12.5%) to "elections" (38.1% => 65%), and that references to "corruption" increased from 7.8% (in 1997) to 22.5% (in 1999), raising hypotheses that the separation was a cause or a consequence for these changes.

In every discovery process, the results from the mining task must be useful for some purpose. Here, we can state that these results may be used to establish political strategies. Examining the distributions of themes, one can evaluate how press is viewing (or manipulating) the events concerning the mayor. For example, "crimes" (32.6%) and "corruption" (10%) are so or more frequent than "education" (26.2%) and "investments" (10.6%). This may help mayor to take care about his declarations and actions or to make a marketing strategy to spread his work. By other side, when looking at the discovered rules, the mayor may decide to stay away from some politicians to avoid having his name associated to "drug traffic" or "loans", for example.

3.2 Competitive Intelligence Experiment

The second experiment had as main goal to compare Text Mining (TM) tools, examining the techniques used and the benefits cited by the vendors of these tools. Another goal was to relate techniques and benefits, in order to discover which techniques to use when needing a certain benefit.

Nine tools were selected using the Copernic meta-search engine (www.copernic.com) and the expression "text mining" as query. Texts about the tools were extracted from the linked web pages (initial and subsequent pages, only those telling about the tool). Which specific tools and URLs were used is not important in this discussion.

Concepts were defined as explained early, resulting in 13 concepts about techniques and 11 concepts corresponding to benefits. Techniques do not cover all the existing or possible ones, but were selected according to a previous survey (*summarization, extraction, key-concept listing, clustering, classification, retrieval, filtering, visualization, hypertext navigation, indexing, NL processing, correlation, sampling)*. The 11 benefits (*ease, support, automation, flexibility, analysis, quickness, completeness, consistency, efficiency, accuracy, conciseness)* were defined to cover all words present in the collection which can be semantically related to benefits. The most interesting patterns are presented below according to the mining technique used.

Concept distributions (key-concept listing technique)

a) the most used techniques are "key-concept listing" and "retrieval" (appearing in all cases) and the least used is "clustering";

b) two tools use all the defined techniques;

c) the most cited benefits were "support" and "ease";

d) there is a tool that alleged 10 benefits.

Associative rules (association technique)

When comparing techniques and benefits, 550 rules were found, from which 281 had confidence equals or greater than 80%. Eight rules had 100% of confidence and 90% of support. Examining these, we noted that "classification ⇔ automation" and "classification ⇔ accuracy" (the sign ⇔ means "if and only if"), perhaps indicating some inherent relation between them.

As in the previous experiment, the discovered knowledge needs to be useful for some purpose. One who intends to create a TM tool can analyze the trends above. If wanting to make something new, he/she must use the clustering technique. If wanting to compete with the existing tools, the most common techniques must be implemented. By other side, marketing strategies over benefits may be also established, using the most cited. And someone who wants to implement a tool to offer a certain benefit should look at the rules comparing benefits versus techniques.

4. DISCUSSION

The discovered knowledge must be interpreted within the context associated to the concept definition. For example, the concept "*corruption*" may be presented in a

situation where the cited mayor was involved in a possible crime or in a situation where the mayor reported a corruption case (for example). Besides that, the findings are relative to the collection. If the texts of the collection are not representative of the real world, we cannot assume that the rules will hold in any situation in a real situation. For example, in the experiments presented in this paper, the results of the KDT process hold only in those collections. Therefore, we cannot state that the same rules apply to the real world, once information present in the texts may be biased by the authors' style, interest, etc.

Given that words in a concept contribute with some weight to the presence of this concept in a text, the decision whether a concept is present or not depends directly on the chosen threshold, but indirectly on the words defined in the concept and their associated weight. We believe that the threshold can cut off undesirable results but may lead to mistakes. One alternative is to use a standard value, proved to be efficient. A future work will evaluate this possibility. Another alternative is to set the threshold using a training sample of the collection, examining errors and hits.

A problem was perceived when a word of a concept appeared in a negative phrase. For example, the expression "*unlike clustering*" erroneously leads to assume that the concept "*clustering*" (a technique in the Text Mining experiment) is present. This is the main problem with Rocchio's and Bayes algorithms, according to [9], once they analyze the whole context (in the entire text) but do not consider the local context (inside a phrase). One simple solution may be the use of negative terms in the concept description, as discussed by [40]. However, texts must be analyzed in the original format or internally represented in other way, perhaps using an ordered list of terms, as proposed by [9]. We know that ambiguity problems can be solved with NLP techniques, however, such algorithms are complex and time consuming, what goes against our initial goals. Our initial solution for this problem is the use of negative values in a concept definition. So, negative words must be selected and included in a concept description with a negative weight. Some experiments have shown that some false hits can be cut off. However, we are still dealing with the whole context and this may still cause false hits and thus more complex methods are necessary (we discuss our directions to solve this problem in the next section).

We also observed problems with ambiguous words. For example, in the TM experiment, the word "*class*" could describe a classification or a clustering technique, what could lead to mistakes. Our suggestion is that the user must examine phrases of a few texts (a sample of the collection) and reduce the weight assigned to this kind of words or eliminate them from the descriptions. We are studying the way in which weights are assigned to terms in a concept description. One alternative is to define them automatically through a training sample, using a supervised method. In an ongoing work, we have implemented different supervised methods for establishing word weights. Initial results suggest that words that appear in all texts within a training set must have a greater degree of importance. Also we are analyzing features such as types of terms (proper nouns, adjectives), sample size and hierarchies of concepts.

In order to compare our approach with other methods to define concepts, we analyzed the Latent Semantic Indexing (LSI), an automatic method that has achieved good results in categorization and information retrieval [12] [13]. The LSI is useful to find relations between terms, where human effort does not bring good results [13]. Thus, the synonymy problem can be solved. However, there are doubts that polysemy can be solved [38]. Deerwester and partners [12] say there is a "partial solution" due to the context analysis, but the consequence may be false hits, like the above example for "*class*". This is because LSI needs a good sample of texts for training, like the most automatic methods. In LSI approach, the sample must be "pure" (each text must be associated to only one category) and "separable" (with a low proportion of terms common to more than one category) [38]. The problem is that good samples (positive and representative cases) are difficult to find, especially in the Web and under the restriction of having only one category per document. Besides that, there is the additional effort (probably by humans) to evaluate the training set and this may also introduce errors.

Even when good samples are available for automatic methods, there is the possibility of negative words being included in a definition, as in the "*unlike*" example, or lexical variations being not considered. This happens because most automatic methods apply statistical techniques and do not use semantic knowledge about the domain. Dumais [13] has conducted experiments using two different methods for acquiring definitions: one extracts words from natural language sentences describing categories and another uses training sets. Although the latter was best on average, there were 14 categories among 50 (28%) for which the former was better. This brings to discussion the relative efficiency of automatic methods.

One problem specific of the LSI approach is that it cuts off infrequent words. Although Yang [48] has showed that this may bring precision improvement in some cases, there is no prove that this will hold in whatever collection and with whatever training sample. Again we rely on good samples. In Yang's experiments, training sets were created with 20, 25 and 50% of the entire collection. Besides, there is the possibility of important infrequent terms (like lexical variations) being ignored and this may cause a great difference when the collection is composed of short texts.

Due to these problems, we chose to use a manual task supported by automatic tools and by existing vocabularies. Automatic methods can help user to find terms related to categories, lexical variations, local synonyms, frequencies and relations between terms. However, human intervention is important to solve ambiguities and to minimize errors. An important step is to examine samples of false hits (texts assigned to wrong categories), looking for terms that lead to errors, in order to refine the concept definitions. In this step, software tools may bring great benefits. Our suggestion is that the final decision should be responsibility of the user, working as a filter.

To evaluate our categorization method, we carried out formal experiments. The first one followed the idea of Goebel and Gruenwald [22], who have done benchmarking of KDD tools. We chose 5 tools from the original paper and defined 13 concepts, 8 relative to tasks and 5 relative to methods (tasks and methods are features evaluated in the original paper). The concepts were described selecting significant terms used in the paper to define each feature. We then gathered texts extracted from the Web pages referred by the paper. As these texts could have different information from those used by the authors in the original benchmarking, we used experts to decide the correct assignments (tools X features). Using Lewis' measures [27], we got the following results:

- microaveraging precision = 0.59;
- macroaveraging precision = 0.54;
- microaveraging recall = 0.95;
- macroaveraging recall = 0.86;
- fallout = 0.62.

Precision and fallout were not good, although in two concepts/categories we found high precision values (1 and 0.83). So, we carried out another experiment, this time performing classification with more accuracy. Some samples of false hits were analyzed and words that led to them (negative terms and ambiguous words) were eliminated from the definitions. With an adjustment in only one concept, the macroaveraging rates increased from 0.54 to 0.61 in precision and from 0.86 to 0.97 in recall (improvement of 13%). A second round was made, redoing all definitions with more accuracy (examining false hits). The task took about 10 minutes and the final results were:

- microaveraging precision = 0.65;
- macroaveraging precision = 0.69;
- microaveraging recall = 0.89;
- macroaveraging recall = 0.93;
- fallout = 0.28.

There was an increase of 10% in microaveraging precision and of 27% in macroaveraging precision. The fallout rate decreased 45%. Recall rates did not have great variance (-6% and +8%). These results show that categorization quality may be improved by refining more the concept definitions and that human intervention may eliminate precision failures.

To show that is possible to achieve a high precision rate, we performed an experiment with a bigger collection. The collection was composed of 100 texts corresponding to medical records of a psychiatric clinic, each one describing the admission interview of a different patient and written by a physician. We performed the categorization comparing each text (between 1 and 4 Kbytes) against 8 concepts. Concepts were manually defined by two experts of the domain (assistant physicians) along 2 months (not full time). The experts examined samples of texts aided by software tools and used Webster's and technical dictionaries to find synonyms. We estimate that the overall time took about 30 hours. Using as threshold the minor value higher than zero, we found an average error (wrong texts inside a concept) equal to 10.8%. Using Lewis' measures for precision [27], we got a microaveraging precision of 0.915 and a macroaveraging precision of 0.891. This reports an increase in the precision rates in comparison to the previous experiments, showing that it is possible to achieve high quality by generating better concept descriptions.

Although the average degrees were good, a special attention must be given to measures in each category (concept). For example, in the medical collection evaluation, there was a concept that got a degree of errors equal to 32.5% while other concepts achieved zero error. The ideal situation is to obtain similar degrees in all categories (concepts) to avoid distorted conclusions.

5. CONCLUDING REMARKS

The paper presented an approach to perform knowledge discovery in textual collections. The process is based on concepts instead of words or attribute values, leading to more real findings, low cost processes and minimizing the *vocabulary problem*. The approach allows users to easily find interesting ideas, ideologies, trends and intentions present in texts, then being useful in sociological studies, discourse analysis, political marketing, competitive intelligence and so on.

An observed advantage is the reduced effort to define and identify concepts in texts, comparing to traditional NLP. The latter is very expensive because it performs complete analysis of a text [40] (using natural

language processing) and because it requires large amounts of formally codified knowledge [25] (knowledge models and extraction rules). By other side, our approach uses simple algorithms and structures, helping people to find interesting patterns in a quick way. In the political experiment for example, classification and categorization did not take more than 40 minutes, remembering that 104 concepts were generated and compared against 358 texts in a Pentium II 400 MHz with 64 Mbytes of RAM. The mining task took about 2 minutes. Chinchor and partners [8] comment that the cost (effort) to adapt MUC-3 systems (classification task) to a new domain was of 10 to 11 man/month per system. It is important to say that the categorization algorithm does not work to extract attribute values, like IE systems, but only works to extract concepts when it is possible to infer them analyzing the presence of words in the whole text. In other situations, the approach should be adapted with other methods for categorization and classification.

We believe that the approach is better suited to interactive discoveries, because user does not need to expend too much effort to define concepts, and does not need to wait a long time for the categorization. Besides that, concept definitions may be refined at execution time. Thus, users with *ad hoc* needs (with ill-defined goals) can perform discovery without having to create formal rules or definitions as ontologies, *thesauri* and IE models. With little knowledge about the domain or even with help of software tools and dictionaries (like a Webster's or a technical one) it is possible to define and refine interesting concepts for a specific application or goal.

Although the problems discussed in the previous section, we believe that the discovery process may have quality if the categorization can be controlled. Based on the evaluations discussed early, we conclude that high quality can be achieved if the concepts are well defined (or refined). A good definition may be obtained if there are available experts, time and predefined vocabularies. Refinements are important and may be done by human intervention through analyzing false hits. The quality level depends on how much effort and resources the user wants or has to expend in the classification task. By other side, we can imagine that concept descriptions may evolve to a more accurate model as users become more familiar with the language used in the documents. McCarthy [34] states that approximate concepts may be refined by learning more or by defining more. But it is necessary to say that the improvement will come only to the specific domain and under the application goal as established by the user. We cannot expect that one specific model (for example, the same set of concept descriptions) will always achieve good results, because the Web is very dynamic and chaotic, as posed by [14]. However, even when concepts are ill defined (there is a lack of a precise definition), McCarthy [34] defends that this does not obstruct us to use and reason about concepts. All concepts are approximate, but they are precise for a certain purpose. So, the results from the KDT process are useful for analyzing trends and do not have compromise with the rigor of a scientific method. The discovered knowledge must be interpreted under this viewpoint.

Another remark is that categorization and classification tasks deserve more attention, especially because they may bias the KDT process. Thus, we are studying other algorithms within the same goals (simplicity, efficiency, ease to use, low cost in time and effort). An ongoing work is the implementation of a classification model using pairs of words and negative words and the implementation of a categorization algorithm that analyzes individual phrases (local context). So, concepts are described by a set of simple rules, each of these composed by positive and negative words. If a phrase has all positive words and no negative word, the concept is assumed to be present in the phrase. An overall computation determines how much the concept is referenced in the whole text. This degree will be used in a further mining task. Initial experiments showed us that some extraction errors (false hits) could be solved. A formal evaluation will be performed to compare the efficiency of the two algorithms.

6. ACKNOWLEDGMENTS

This work is partially supported by: CNPq (Brazilian Council for Scientific and Technological Development), ULBRA (Lutheran University of Brazil) and UCPEL (Catholic University of Pelotas). We would like to thank our advisor Prof. Dr. José Palazzo Moreira de Oliveira and the anonymous referee for the valuable suggestions.

7. REFERENCES

[1] Apté, Chidanand; Damerau, Fred; Weiss, Sholom M. Automated learning of decision rules for text categorization. ACM Transactions on Information Systems, v.12, n.3, July 1994.

[2] Bates, M. J. Subject access in online catalogs: a design model. Journal of the American Society for Information Science, v.37, n.6, November 1986.

[3] Buckley, Chris; Salton, Gerard; Allan, James. The effect of adding relevance information in a relevance feedback environment. In: VII International ACM-SIGIR Conference on Research and Development in Information Retrieval. London: Springer-Verlag. 1994.

[4] Chakrabarti, Soumen. Data mining for hypertext: a tutorial survey. ACM SIGKDD Explorations, v.1, n.2, January 2000.

[5] Chen, Hsinchum. The vocabulary problem in collaboration. IEEE Computer, special issue on CSCW, v.27, n.5, May 1994. Online at http://ai.bpa.arizona.edu/papers/ cscw94/cscw94.html

[6] Chen, Hsinchum e al. Automatic concept classification of text from electronic meetings. Communications of the ACM, v.37, n.10, October 1994. Online at http://ai.bpa.arizona.edu/papers/ ebs92/ebs92.html

[7] Chen, Hsinchum et al. A concept space approach to addressing the vocabulary problem in scientific information retrieval: an experiment on the worm community system. Journal of the American Society for Information Science, v.47, n.8, August 1996. Online at http://ai.bpa.arizona.edu/papers/wcs96/wcs96.html

[8] Chinchor, Nancy; Hirschman, Lynette; Lewis, David D. Evaluating message understanding systems: an analysis of the third message understanding conference (MUC-3). Computational Linguistics, v.19, n.3, September 1993.

[9] Cohen, William W. and Singer, Yoram. Context-sensitive learning methods for text categorization. In: International ACM-SIGIR Conference on Research and Development in Information Retrieval SIGIR-96. 1996. Online at http://www.research.att.com/~wcohen/ index.html

[10] Cowie, Jim and Lehnert, Wendy. Information extraction. Communications of the ACM, v.39, n.1, January 1996.

[11] Croft, W. Bruce. Machine learning and information retrieval. In: COLT '95 Conference. Lake Tahoe, July 1995. (invited talk) Online at http://www.ee.umd.edu/medlab/filter/

[12] Deerwester, Scott et alli. Indexing by latent semantic analysis. Journal of the Society for Information Science, v.41, n.6, 1990.

[13] Dumais, Susan T. Combining evidence for effective information filtering. In: AAAI Spring Symposium on Machine Learning and Information Retrieval, Tech Report SS-96-07, AAAI Press, March 1996.

[14] Etzioni, Oren. The world-wide web: quagmire or gold mine ? Communications of the ACM, v.39, n.11, November 1996.

[15] Feldman, Ronen and Dagan, Ido. Knowledge discovery in textual databases (KDT). In: 1st International Conference on Knowledge Discovery (KDD-95). Montreal, August 1995.

[16] Feldman, Ronen and Hirsh, Haym. Exploiting background information in knowledge discovery from text. Journal of Intelligent Information Systems, v.9, n.1, July/August de 1997.

[17] Feldman, Ronen and Dagan, Ido. Mining text using keyword distributions. Journal of Intelligent Information Systems, v.10, n.3, 1998.

[18] Fisher, Douglas H. Knowledge acquisition via incremental conceptual clustering. Machine Learning, v. 2, pp.139-172. 1987. Reprinted in Shavlik & Dietterich (eds.), Readings in Machine Learning, section 3.2.1.

[19] Frawley, W. J.; Piatetsky-Shapiro, G.; Matheus, C. J. Knowledge discovery in databases: an overview. In: Piatetsky-Shapiro, G.; Frawley, W. J. (eds.). Knowledge discovery in databases. MIT Press. 1991.

[20] Furnas, G. W. et al. The vocabulary problem in human-system communication. Communications of the ACM, v.30, n.11, November 1987.

[21] Garofalakis, Minos N. et al. Data mining and the web: past, present and future. In: ACM Workshop on Information and Data Management, Kansas City, 1999.

[22] Goebel, Michael and Gruenwald, Le. A survey of data mining and knowledge discovery software tools. ACM SIGKDD Explorations, v.1, n.1, June 1999.

[23] Gulla, Jon A. et alli. An abductive, linguistic approach to model retrieval. Data & Knowledge Engineering, v.23, n.1, June 1997.

[24] Iivnen, Mirja. Searches and searches: differences between the most and least consistent searches. In: International ACM-SIGIR Conference on Research and Development in Information Retrieval SIGIR'95. Washington: ACM PRESS, 1995.

[25] Knight, Kevin. Mining online text. Communications of the ACM, v.42, n.11, November 1999.

[26] Lagus, K. and Kaski, S. Keyword selection method for characterizing text document maps. In: Ninth International Conference on Artificial Neural Networks – ICANN'99, volume 1, pages 371-376, IEE, London (1999). Online at http://websom.hut.fi/websom/doc/ publications.html

[27] Lewis, David D. Evaluating text categorization. Proceedings of the Speech and Natural Language Workshop, Asilomar, February 1991. Online at http://www.research.att.com/ ~lewis

[28] Lewis, David D. and Hayes, Philip J. Guest editorial. ACM Transactions on Information Systems, v.12, n.3, July 1994.

[29] Liddy, Elizabeth D.; Paik, Woojin; Yu, Edmund S. Text categorization for multiple users based on semantic features from a machine-readable dictionary. ACM Transactions on Information Systems, v.12, n.3, July 1994.

[30] Lin, Chung-hsin and Chen, Hsinchun. An automatic indexing and neural network approach to concept retrieval and classification of multilingual (Chinese-English) documents. IEEE Transactions on Systems, Man and Cybernetics, v. 26, n.1, February 1996. Online at http://ai.bpa.arizona.edu/papers/chinese93/chinese93.html

[31] Lin, Shian-Hua et al. Extracting classification knowledge of Internet documents with mining term associations: a semantic approach. In: International ACM-SIGIR Conference on Research and Development in Information Retrieval (SIGIR-98). 1998.

[32] Mannila, Heikki, Theoretical frameworks for data mining. ACM SIGKDD Explorations, v.1, n.2, January 2000.

[33] Mattox, David; Seligman, Len; Smith, Ken. Rapper: a wrapper generator with linguistic knowledge. In: ACM Workshop on Information and Data Management, Kansas City, 1999.

[34] McCarthy, John. Approximate objects and approximate theories. In: Seventh International Conference on Principles of Knowledge Representation and Reasoning (KR2000). April 2000. Online at http://www-formal.stanford.edu/jmc

[35] Miller, George A. WordNet: A lexical database for English. Communications of the ACM, v.38, n.11, 1995

[36] Morris, Charles W. Foundations of the theory of signs. Rio de Janeiro, Eldorado Tijuca. 1976. (in Portuguese)

[37] Nakanishi, H.; Turksen, I. B.; Sugeno, M. A review and comparison of six reasoning methods. Fuzzy Sets and Systems, 57, 1993.

[38] Papadimitriou, Christos H. et alli. Latent Semantic Indexing: a probabilistic analysis. In: Seventeenth ACM SIGACT-SIGMOD-SIGART International Conference on Management of Data and Symposium on Principles of Database Systems (PODS). Seattle, June 1998..

[39] Ragas, Hein and Koster, Cornelis H. A. Four text classification algorithms compared on a Dutch corpus. In: International ACM-SIGIR Conference on Research and Development in Information Retrieval (SIGIR'98). Melbourne, 1998.

[40] Riloff, Ellen and Lehnert, Wendy. Information extraction as a basis for high-precision text classification. ACM Transactions on Information Systems, v.12, n.3, July 1994.

[41] Rocchio, J. J. Document retrieval systems - optimization and evaluation. Ph.D. Thesis, Harvard University, Report ISR-10 to National Science Foundation, Harvard Computation Laboratory. 1966.

[42] Salton, G. and McGill, M. J. Introduction to modern information retrieval. McGraw-Hill, 1983.

[43] Soderland, Stephen. Learning to extract text-based information from the world wide web. In: 3rd International Conference on Knowledge Discovery and Data Mining (KDD-97). 1997.

[44] Sowa, John F. Knowledge representation: logical, philosophical, and computational foundations. Brooks/Cole Publishing Co., Pacific Grove, CA, 2000.

[45] Sparck-Jones, Karen. Assumptions and issues in text-based retrieval. In Jacobs, Paul S. (ed.) Text-based intelligent systems: current research and practice in information extraction and retrieval. New Jersey: Lawrence Erlbaum, 1992.

[46] Wiener, Erik D. et al. A neural network approach to topic spotting. In: 4th Annual Symposium on Document Analysis and Information Retrieval (SDAIR'95), Las Vegas, 1995. Online at http://www.stern.nyu.edu/~aweigend/Research/Papers/TextCategorization

[47] Yang, Yiming and Chute, Christopher G. An example-based mapping method for text categorization and retrieval. ACM Transactions on Information Systems, v.12, n.3, July 1994.

[48] Yang, Yiming. Noise reduction in a statistical approach to text categorization. In: ACM-SIGIR Conference on Research and Development in Information Retrieval (SIGIR'95) Seattle, 1995.

[49] Zadeh, Lotfi A. Outline of a new approach to the analysis of complex systems and decision processes. IEEE Transactions on Systems, Man and Cybernetics, v. SMC-3, n.1, January 1973.

About the authors:

Stanley Loh is an assistant teacher in the Lutheran University of Brazil (ULBRA) and in the Catholic University of Pelotas (UCPEL). He obtained a Master degree in Computer Science in 1991 at the Federal University of Rio Grande do Sul (UFRGS), where he is now a Ph.D. candidate. His researches are about knowledge discovery, text mining, information retrieval and tools for competitive intelligence.

Leandro Krug Wives currently works with computational methods and tools for competitive intelligence (business intelligence). He obtained his Master degree in Computer Science in 1999 at the Federal University of Rio Grande do

Sul (UFRGS). Now he is continuing his work as Ph.D. candidate in the same university. His main interests are information retrieval, text mining and competitive intelligence.

José Palazzo Moreira de Oliveira is a full professor of Computer Science at Federal University of Rio Grande do Sul (UFRGS). He has a doctor degree in Computer Science from Institut National Politechnique - IMAG (1984), Grenoble, France, a M.Sc. degree in Computer Science from PPGC-UFRGS (1976) and Electronic Engineer, EE-UFRGS (1968). His research interests include information systems and industrial conceptual modeling, tele-education, computer based design support tools, temporal databases, applications of database technology and distributed systems. He has published about 130 papers, being advisor of 9 Ph.D. and 35 M.Sc. students.