KNOWLEDGE DISCOVERY IN TEXTUAL DOCUMENTATION: QUALITATIVE AND QUANTITATIVE ANALYSES

STANLEY LOH

sloh@zaz.com.br

Post-Grade in Computer Science, Federal Univ. of Rio Grande do Sul, Porto Alegre, Brazil

Computer Science Department, Catholic University of Pelotas, Pelotas, Brazil

Computer Science Department, Lutheran University of Brazil, Canoas, Brazil

JOSÉ PALAZZO M. de OLIVEIRA

palazzo@inf.ufrgs.br

Institute of Computer Science, Federal Univ. of Rio Grande do Sul, Porto Alegre, Brazil

FÁBIO LEITE GASTAL

flgastal@zaz.com.br

Olivé Leite Hospital, Pelotas, Brazil

Medicine Department, Catholic University of Pelotas, Pelotas, Brazil

This paper presents an approach for performing knowledge discovery in texts through qualitative and quantitative analyses of high-level textual characteristics. Instead of applying mining techniques on attribute values, terms or keywords extracted from texts, the discovery process works over concepts identified in texts. Concepts represent real world events and objects, and they help the user to understand ideas, trends, thoughts, opinions and intentions present in texts. The approach combines a quasi-automatic categorisation task (for qualitative analysis) with a mining process (for quantitative analysis). The goal is to find new and useful knowledge inside a textual collection through the use of mining techniques applied over concepts (representing text content). In this paper, an application of the approach over medical records of a Psychiatric Hospital is presented. The approach helps physicians to extract knowledge about patients and diseases. This knowledge may be used for epidemiological studies, for training professionals and it may be also used to support physicians to diagnose and evaluate diseases.

1. Introduction

With the growing use of digital resources, people and organisations have stored a great volume of documents. This digital documentation has hidden knowledge, implicit in relations within and among documents (Davies, 1989). In most cases, people and organisations have difficulty to analyse this massive documentation in order to extract new and useful information to improve the knowledge about the domain.

The novel area called *Knowledge Discovery in Texts* (KDT) has emerged to help people to extract knowledge from textual documents, through the application of techniques from *Knowledge Discovery in Databases* (KDD) over texts (Feldman & Dagan, 1995). KDD is the "*nontrivial extraction of implicit, previously unknown, and potentially useful information from given data*" (Frawley et al., 1991) and it has obtained success applying statistical mining techniques in large databases. However, researches on KDD deals only with structured data (tables, records and fields/attributes). By other side, texts have information coded in natural language sentences (free and unstructured phrases). As a result, information may appear in different styles or formats, making difficult to discover it. The goal of KDT is to extract information from texts and explore the results in order to find new and interesting knowledge. Knowledge is defined as useful information; people receive large amounts of information but only part of this set becomes knowledge since not all information is employed in the discovery process.

This paper presents an approach for knowledge discovery in texts through qualitative and quantitative analyses of high-level textual characteristics exploring the content present in a textual collection. In this approach, instead of applying mining techniques on attribute values, terms or keywords extracted from texts, the discovery process works over concepts identified in texts. Concepts represent real world events and objects, and they help the user to understand ideas, trends, thoughts, opinions and intentions present in texts. The approach combines a quasi-automatic categorisation task (qualitative analysis) with a mining process (quantitative analysis). Categorisation identifies concepts in texts and mining discovers patterns by analysing and relating concept distributions in a collection. The goal is to find new and useful knowledge inside a collection. The discovered knowledge may be used for decision support and evaluation, training of workers and analysis of domain characteristics.

In this paper, an application of KDT concerning medical records of a Psychiatric Hospital is presented. The approach helps physicians to extract knowledge about patients and diseases. This

knowledge may be used for epidemiological studies, for training professionals and it may be also used to support physicians to diagnose and evaluate diseases.

Section two analyses related works and explains some problems with the existing approaches. Section three describes the proposed approach in a general view. Section four presents the results of applying the approach in a textual documentation of a Psychiatric Hospital. Section five discusses the quality of the approach and section six presents concluding remarks and future works.

2. Related Work

Lin et al. (1998) use terms automatically extracted from the text to characterise documents and to find associations or co-relations. The most frequent terms are assigned as keywords (attributes). However, when analysing words, problems arise due to the *vocabulary problem*. The language use may cause semantic mistakes due to synonymy (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-synonymy (words related to the same subject, object or event, like "bomb" and "terrorist attack") (Chen, 1994), (Chen et al., 1997) and (Furnas et al., 1987). For example, a murder may be described with terms like "murder" or "homicide". If analysing only the terms, the discovery process may be misled by semantic gaps.

Other interesting approach for KDT is to apply KDD techniques after the use of Information Extraction (IE) techniques, which transform information present in texts in attribute values of a structured database (Cowie & Lehnert, 1996). However, IE systems use complex rules, usually based on natural language processing techniques. This process requires a complex computational algorithm and a great human effort of knowledge engineering to understand how information is coded in natural language (Chinchor et al., 1993) (Gaizauskas & Wilks, 1998).

Feldman & Dagan (1995, 1998) face the KDT problem applying mining techniques over keywords that are assigned to texts (as attributes). These mining techniques use statistical analysis to discover association rules and interesting patterns on keyword distributions and associations. In the cited works, keywords should be previously assigned to texts, either by human or by automatic tasks. When using human assigned keywords, in general only part of all themes present in a text is available for analysis because the human effort concentrates on meaningful aspects. Furthermore, there is the extra work of reading and understanding the text. Automatic tasks correspond to text categorisation and are more suitable for that problem.

However, most text categorisation works rely on a single class method, that is, they find "the class" to which the text belongs. Methods that find more than one class should be employed to get a wider and more significant categorisation. This is a mandatory requirement in medical systems as the symptom complexity leads to a more complex classification process.

Wiener et al. (1995) use neural networks to extract topics from texts (many classes). One problem of this approach is that it is only used for text categorisation (they call it "topic spotting"); no quantitative analysis is done. Another problem is that the extracted knowledge is not used for later analyses or processes, as training, decision support and evaluation.

3. The Approach for KDT

The proposed approach for knowledge discovery analyses high-level characteristics of texts, allowing qualitative and quantitative analyses over the content of a textual collection. Instead of applying mining techniques on attribute values, terms or keywords extracted from texts, the discovery process works over concepts identified in the texts. Concepts represent real world events and objects, and they help the user to explore, examine and understand the contents (ideas, ideologies, trends, thoughts, opinions and intentions) of talks, texts, documents, books, messages, etc. Chen et al. (1994), 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 & Chen (1996) 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 the present approach, the categorisation main advantage is to minimise the vocabulary problem.

The proposed KDT approach combines a quasi-automatic categorisation task with a mining task. Categorisation identifies concepts in texts (qualitative analysis) and mining discovers patterns by analysing and relating concept distributions in a collection (quantitative analysis).

3.1 The categorisation process

The goal of the categorisation is to identify concepts present in texts. However, documents do not have concepts explicitly stated, but instead they are composed of words that represent the concepts. As concepts are expressed by language structures (words and grammars), it is possible to identify concepts in texts analysing phrases (Sowa, 2000). However, we need some straightforward

algorithm to accomplish this purpose. The method used in our approach identifies concepts analysing individual phrases of a text. The goal is to verify whether a concept is mentioned in a phrase.

Consequently each phrase of a text is compared against rules that define a concept (and against all concepts defined for a domain). Rules combine positive and negative words. To a concept be present in a phrase, all positive words must be present and none negative word may be present. If one of the concept rules is true, then the concept is present in the phrase and consequently is also present in the text. For example, in a medical domain, "headache" (a symptom) may be defined as a concept using the following rules (negative words have a '-' before):

(i) headache –deny –denies

(ii) head pain –deny –denies

If the concept is present more than once in a text, the total counting is used to define an associative degree between the text and the concept, indicating how much a concept is referred by a text.

The definition of the concepts may be generated in different ways. The proposed approach uses a combination of automatic tools and human decision. Automatic tools help people to have an insight of the language used in the texts (different terms and meanings). Humans, in the other hand, augment this vocabulary using Webster-like dictionaries or technical dictionaries. Software tools can also be useful to analyse textual samples in order to verify if the defined rules work correctly. False hits may help in defining negative words. The final decision about the rules in each concept definition should be responsibility of humans. In other paper more details about some tactics for defining concepts are discussed (Loh et al., 2000).

3.2 The mining process

The approach here described uses distribution analyses to discover interesting patterns. The first technique used is the key-concept listing, which analyses concept distributions over the collection. A software tool counts the number of texts where each concept is present, generating a vector of concepts and their distributions inside the collection (called centroid). Different centroids can be generated for different collections or for parts of a unique collection (sub-collections). This technique allows finding what dominant themes exist in a collection, in a sub-collection or in a single text. In addition, we can compare one centroid to another (between sub-collections), to find common themes or variations between sub-collections. Another possible usage is to find differences

between sub-collections (concepts present in only one text). Feldman & Dagan (1998) suggest the exam of 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 Lin et al. (1998), *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. Confidence works like the conditional probability (*if* X *is present, so there is a certain probability of* Y *being present too*). This allows predicting the presence of a concept associated to the presence of another concept. Complex rules may be discovered with human intervention. As a result, 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.

The choice for these two techniques is due to their simplicity and extensive use in Data Mining approaches (KDD). One hypothesis is that other different techniques can be used over the concepts, after the categorisation task.

4. Example of an Application

Some experiments have being carried out on a collection of medical records from a psychiatric hospital. This domain has special characteristics, as the diagnosis process is more complex than in other medical specialities. Symptoms and signals may be present in different diseases and there are not syndrome definitions, relating symptoms and signals to a specific disease. Other problem is that symptoms and signals may be present in a moment and disappear in other. Besides that, some characteristics may be more predominant than others in a period and a different situation may occur in another time.

The goal of the experiments is to discover knowledge about this domain, so that the results may be used to help physicians in the diagnosis process, to evaluate diagnosis decisions and to

qualify students or trainees. From this point of view, the discovery method, described early, has two advantages over others:

1) the analyses are performed on concepts present in the texts instead of on individual words; thus patient symptoms, signals and other characteristics can be analysed (qualitative analysis);
2) the knowledge extracted from the concept distribution analysis (quantitative analysis) may be used to
    a) automatically classify patients in diseases (diagnosis): in the psychiatric domain, inductive decision trees are not well suited, since characteristics may be present in more than one class; consequently methods like ID3 and C4.5 are not indicated in this case; see Ingargiola (1996) for more details on these algorithms;
    b) understand how the process was developed: unlike neural networks, the rules used to identify the class are available to explain why a certain class was associated to a test case.

In this application, the first medical record of patients was used, created in the patient admission. Physicians generated the texts after interviewing the patient and his/her relatives. These records include the patient history and do not have explicitly stated the final diagnosis. Information about the patient concerns the diary activities, social and familiar behaviour and past medical history if readmitted. The records also contain symptoms and signals identified by the physician during the interview.

Two different collections were used. The first one was composed of 200 texts, each one corresponding to only one patient (remembering that it could be a readmission). This collection was used for the training process, to discover knowledge about the domain. The second collection had 200 different texts and was used for the test process, to evaluate the discovered knowledge quality. Some texts in the two collections could correspond to the same patient. Each collection corresponds to admissions made during a two months period.

The texts were classified in one of four major classes, corresponding to diseases of the International Classification of Diseases, 10^{th} revision - ICD-10 (BCDC, 1993). Physicians in a real diagnosis process previously determined the class (disease). The classes were:

a) ***organic*** mental disturbances (due to brain damage), including codes F00 to F09 of the ICD-10;

b) mental and conduct disturbances due to psychoactive ***substances***, including codes F10 to F19;

c) ***schizophrenia***, schizoid disorders and delirious disturbances, including codes F20 to F29;

d) ***affective*** and mood disturbances, including codes F30 to F39.

The first collection (for training) was composed of: 27 texts from "*affective*" (13.5%), 103 texts from "*schizophrenia*" (51.5%), 18 texts from "*organic*" (9%) and 52 texts from "*substances*" (26%). The second collection (for test) was very similar (*affective* – 12.4%, *schizophrenia* – 52.7%, *organic* – 8.4% and *substances* – 26.3%). All texts ranged from 1 to 4 Kbytes in size.

4.1 Concepts used

The concept definition task selected 65 concepts, corresponding to symptoms, signals and social characteristics (for example, *inappetence, insomnia, aggressiveness, tobacco use, living alone)* or referencing events, persons or objects (for example, *marriage, husband, wife, children, neighbours, knife, weapon, hanging*). The stated goal was to identify references inside the texts, which could be important to characterise the diseases of the patients.

For selecting the concepts, ICD reports and dictionaries from psychiatry were used. Also all words and terms used in texts of the training collection were examined, in order to find important references to events, persons or objects. Software tools were used in this last task.

After that, the rules for each concept were defined through the same process. Additional software tools were used to examine the context of words and terms. Special attention was given to synonyms, which could be identified analysing the documents with help of a Webster's dictionary. These choices are explained by Loh et al (2000). Two professionals helped in the process, taking approximately 30 hours at all, during a 2 months period. The final decision is due to these professionals.

Below some examples of the used concepts and their definitions are showed (each rule is identified by a roman number; the symbol '$' indicates a radical and '–' indicates a negative word):

- "*alcoholism*":

(i) alcohol$ (ii) ethilic (iii) drink (iv) drunk (v) drank; etc.

- "*inappetence*":

(i) eat not much (ii) feed badly (iii) fed not much; etc.

- "*homicide*":

(i) kill$ –himself –herself (ii) homicid$; etc.

- "*relatives*":

(i) mother (ii) father (iii) brother$ (iv) sister$ (v) uncle; etc.

4.2 Analyses

The mining process was oriented to analyse the distribution of the concepts in the whole collection to identify which concepts are the most frequent. Assuming that the collection is a representative sample of all records in the Hospital, we can make predictions about new patients or use this knowledge for epidemiological studies.

In addition, concept distributions for the four classes (representing diseases) were also compared, looking for similar and very different values.

Finally, using additional software tools, the collection was separated by medicine or drug administration. Analysing the concept distributions inside each sub-collection, it was possible to identify which concepts were dominant and thus to infer the symptoms and signals for which the medicines/drugs are indicated.

For the training collection, the time for categorisation (only the identification of concepts in the texts) took about 1 hour and 20 minutes in a Pentium II 400 MHz with 64 Mbytes of RAM (a comparison of 200 texts against 65 concepts). The mining process took about 15 minutes.

4.3 Discovered Knowledge

<u>Analysis of the whole collection</u>

- Most frequent concepts (above 50%): relatives (84.5%), aggressiveness (77%), inappetence (76%), medicines (74.5%), insomnia (71%), thought deficit (70.5%), nervousness (68.5%), attention deficit (54.5%)

- Interesting observations:

a) "*readmission*" = 33.0% (meaning 1/3 of the internal patients are readmitted);

b) 84.5% of patients have *relatives*;

c) "*aggressiveness*" is the most frequent symptom in the collection.

Analysis by sub-collection (disease)

Comparing the concept distributions among the four classes in the training collection, some patterns arose. Only concepts with very different distributions were considered as interesting patterns, since similar distributions do not help in discriminating different classes. No knowledge from experts was used to select interesting patterns.

Then these patterns were tested in the second collection. The assumption is that a pattern that appears in both collections is more probable to be correct. Below, the patterns verified in both collections are presented: (when not indicated, percentages follow the order: *affective, schizophrenia, organic, substances*)

1- all the concepts appear in more than one class, except "*poison*", which only appears in *schizophrenia* but with a small frequency;
2- "*attention deficit*" is more frequent in the *affective* class;
3- "*suicidal*" is more frequent in *affective* (81.5% against 38.8%, 16.7% and 30.8%)
4- "*depression*" appears more in *affective* (74.1% against 11.7%, 11.1% and 25%)
5- *affective* and *substances* are very similar (similar frequencies for "*insomnia*", "*inappetence*", "*nervousness*", "*aggressiveness*"), except for "*alcoholism*" (high in the latter and low in the former) and for "*depression*", "*suicidal*" and "*crying*" (on the contrary)
6- "*autism*" appears only in *schizophrenia* (37.9%) and in *organic* (16.7%)
7- "*alcoholism*" appears more in *substances* (94.2% against 25.9%, 16.5% and 11.1%)
8- "*normal consciousness*" and "*clouding of consciousness*" had similar distributions in *substances* (17.3%) and in *organic* (11.1%), while in *affective* and in *schizophrenia*, "*normal consciousness*" is more frequent than "*clouding of consciousness*" (40.7% x 7.4%; 32% x 13.6%)
9- references to "*death*" are lower in *substances* (17.3% against 40.7%, 35% and 33.3%)
10- "*negativism*" has low frequency in *substances* (17.3% against 29.6%, 38.8% and 38.9%)
11- "*insights*" and "*animals*" (*zoopsia*) do not appear in *organic*
12- "*injuries*" are higher in *organic* (38.9% against 18.5%, 13.6% and 21.2%)
13- "*living alone*" does not appear in *organic* and it is low in the others (14.8%, 8.7% and 3.8%)
14- "*marriage*", "*husband*" and "*wife*" do not appear in *organic*
15- "*puerile*" does not appear in *substances*
16- "*mania*" does not appear in organic and is low in *affective* and *substances* (7.4% and 5.8%)
17- "*dromomania*" does not appear in *organic* and *substances* and is low in *affective* (7.4%)
18- "*trembling*" does not appear in *schizophrenia* and *organic*, is high in *substances* (40.4%) and low in *affective* (7.4%)
19- "*tobacco use*" is not cited in *organic*
20- "*delirium*" does not appear in *affective*
21- concept distributions in "*readmission*" sub-collection (records of readmitted patients) are similar to the whole collection
22- other concepts did not showed an interesting pattern

Associative rules (by diagnosis)

Using a confidence threshold of 80% and a support threshold equals to 40%, one set of associative rules was discovered for each class. The sets were compared to find the common rules (true for all diseases) and the exclusive rules (which appear in only one disease). Following, some associative rules per group are presented:

Common Rules:

attention deficit ➔ relatives
attention deficit ➔ thought deficit
inappetence ➔ relatives
insomnia ➔ relatives
thought deficit ➔ relatives
medicines ➔ relatives

Substances:

aggressiveness ➔ inappetence
thought deficit ➔ inappetence
work/job ➔ inappetence

Schizophrenia:

voices ➔ aggressiveness
persecution ➔ insomnia
voices ➔ insomnia
voices ➔ nervousness
persecution ➔ thought deficit
voices ➔ thought deficit

Organic:

injuries ➔ aggressiveness
injuries ➔ nervousness
negativism ➔ aggressiveness

Affective:

inappetence ➔ suicidal
insomnia ➔ suicidal
thought deficit ➔ suicidal

Analysis of drugs/medicines (an example)

Following, concept distributions for the medicine Dienpax are presented (percentages indicate how frequent is the concept in the set of records where this medicine appears): *inappetence* (91.8%), *aggressiveness* (83.7%), *thought deficit* (78.3%), *nervousness* (75.6%), *insomnia* (64.8%), *alcoholism* (62.1%), *voices* (59.4%).

5. Results Evaluation

It was necessary to evaluate whether the discovered knowledge was true. First the evaluation of the categorisation process was carried out, intending to determine the error rate in identifying concepts inside the texts. If the error was too great, that would mislead the mining process.

Second, the discovered knowledge needed to be validated against the real expertise about the domain. Two approaches were used for this validation: one subjective and other objective. The subjective validation consisted in the presentation of the results to psychiatrists in order to obtain expert feedback. For the objective validation, an automatic system was constructed for determining the disease of test cases representing patients without diagnosis. The discovered knowledge was used in the decision algorithm of this system. The evaluation was to compare the diagnosis indicated by the automatic system against the one predetermined by physicians in the test collection.

5.1 Categorisation Process Evaluation

A sample of 50 texts extracted from the test collection was examined to evaluate the concept identification. For this evaluation, twelve concepts were selected: those more prone to errors (with complex rules). Recall and precision values were calculated using *microaveraging* and *macroaveraging* measures of Lewis (1991). *Microaveraging* considers the whole collection as a unique class and *macroaveraging* first calculates precision and recall inside each class and then extracts the average value for the entire collection.

The results were:

- *microaveraging* precision = 90%
- *microaveraging* recall = 93%
- *macroaveraging* precision = 89%
- *macroaveraging* recall = 92%

An average error of 10% may be considered a good result. The improvement of these rates is possible by analysing samples of texts with false hits (causing low precision) or the disregarded texts (low recall) and then refining the categorisation rules. The results described above were obtained in a final round, after improving concept definitions. In a previous process, the results generated low values, respectively 75%, 87%, 71% and 87%. This proves that it is possible to minimise the error, refining the rules for concept identification.

However, a special attention must be given to errors in each concept. For example, the concept "*depression*" had the worst precision (73%) and the concept "*death*" had the worst recall value (73%). These results put in doubt the extracted knowledge concerning these concepts.

5.2 Discovered Knowledge Evaluation

The subjective evaluation of the discovered knowledge was done presenting the results to two expert physicians in Psychiatry. The response was that the knowledge is very similar to that used in real processes for diagnosis. This feedback was enough to consider reliable the results of this experiment.

A final objective evaluation was carried out. The discovered knowledge was used in an automatic classification system for identifying the disease of the 200 test texts (diagnosis process). Different classification methods were experimented, for example:

a) using as class descriptors the concept distribution of each class;

b) using the least frequent concepts in each class and their distributions as weights;

c) using negative concepts (those that never appear in a class) to discard a diagnosis;

d) using negative concepts with negative weights;

e) using pairs of concepts (according to exclusive associative rules).

The best method, a combination of (b) and (d), achieved the following results:

- *microaveraging* precision = 65%

- *microaveraging* recall = 73%

- *macroaveraging* precision = 61%

- *macroaveraging* recall = 53%

Presenting these results to the same physicians (an average error of 38%), they considered a very good rate, above some human performances.

6. Concluding Remarks

This paper presented an approach for performing knowledge discovery in texts through the qualitative and quantitative analyses of high-level textual characteristics. The analyses are performed based on concepts instead of words.

The process is well suited for analysing textual documentation present in organisations. Qualitative analysis discovers concepts referenced in the documentation and quantitative analysis allows the examination of concept distributions and relations.

The goal is to discover new and useful knowledge through this examination. The results may be used for supporting and evaluating decision processes and for training professionals.

The paper presented the application of the approach in textual documents from a psychiatric hospital. Concepts represent symptoms, signals and social characteristics of patients and were used to identify diseases.

Subjective and objective evaluations were carried out to validate the discovered knowledge. Results proved that the knowledge is reliable and thus the approach has obtained success. Results from the objective evaluation (with the automatic diagnosis system) demonstrate that the approach may be used for constructing decision support systems.

Assuming that the collection used in the experiments is representative of all patients, it is also possible to predict the characteristics of new patients and perform epidemiological studies (for example, to identify geographic causes of diseases). However, the sample could be conditioned to some aspects, for example, seasons and external events. Currently, texts form a unique set and have no time associated. A future work is planned to analyse the concept distributions over different time periods (years, seasons).

In the example application, a few concepts correspond to social characteristics, but other concepts may be generated and analysed. Consequently, other medicine areas may also benefit from this approach.

As physicians considered the discovered patterns similar to the knowledge used in their decision processes, the results of the discovery approach may be used for training students and assistant professionals. Furthermore, the patterns concerning concepts instead of words make the knowledge more clear and understandable.

The psychiatric collection is a special case for applying knowledge discovery. First, because the diagnosis process is very complex and some times, it is normal that physicians disagree about the final decision. Also it is usual to occur later corrections in the diagnosis associated to a patient. In the presented application, the texts correspond to the admission record but the associated diagnosis is the more updated one. That is, more information (not available in the first record) may have been used to make the final decision.

Special attention should be given to some aspects of the approach. First, the presence of a concept inside a text is conditioned to the concept interpretation. For example, when the concept "*alcoholism*" is identified in a text, the correct interpretation is that the concept is cited in the text but the cause may be doubtful. In order to avoid errors, as when the concept is cited because someone in the family uses alcohol, it is necessary to carefully define the rules for each concept.

Regarding the errors in the process, it is possible to minimise the rates, but perhaps never eliminate them at all. However, the error rate can be controlled and the results may be interpreted under a certain degree of reliability.

Other remark is that the language may change along the time (Chen, 1994). The way in which the vocabulary is used may vary according to people and situations. Furthermore, when the own world evolves, concepts and languages also evolve to accommodate the changes. Therefore, the definition of concepts is conditioned by this context and the analyses should be performed under these restrictions.

This work have used two mining techniques (key-concept listing and associative), since they are the most used in knowledge discovery. However, other techniques can be used in the mining process after the extraction of concepts (categorisation task). The next step in this work is to apply the time series technique over textual records describing the evolution of the patient. Analysing the sequence of records, it is possible to find associations between concepts along the time, for example, discovering concepts that appear immediately after some drug administration or some time after a certain symptom registration.

Acknowledgements

This research is partially sponsored by: CNPq (Brazilian Council for Scientific and Technological Development) and CAPES. Medical records were provided by Olivé Leite Psychiatric Hospital (Pelotas, RS, Brazil) and have being produced with research support by FIDEPS (Found of Incentive for the Development of Teaching and Research in Health - Ministry of Health, Brazil).

References

Brazilian Center for Disease Classification. (1993) *International Classification of Diseases and Health Related Problems in Portuguese*. 10th revision. São Paulo: EDUSP (in collaboration with the World Health Organisation).

Chen, H. (1994). The vocabulary problem in collaboration. *IEEE Computer, special issue on CSCW*, 27(5), p.2-10.

Chen, H. et al. (1994). Automatic concept classification of text from electronic meetings. *Communications of the ACM*, 37(10), p.56-73.

Chen, H. et al. (1997). 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*, 48(1), p.17-31.

Chinchor, N. et al. (1993). Evaluating message understanding systems: an analysis of the third message understanding conference (MUC-3). *Computational Linguistics*, 19(3), p.409-449.

Cowie, J. & Lehnert, W. (1996). Information extraction. *Communications of the ACM*, 39(1), p.80-91.

Davies, R. (1989). The creation of new knowledge by information retrieval and classification. *Journal of Documentation*, 45(4), p.273-301.

Feldman, R. & Dagan, I. (1995). Knowledge discovery in textual databases (KDT). In: Fayyad, Usama & Uthurusamy, Ramasamy, eds. *Proceedings of the 1st International Conference on Knowledge Discovery* (KDD-95). Cambridge: AAAI/MIT Press, p.112-117.

Feldman, R. & Dagan, I. (1998). Mining text using keyword distributions. *Journal of Intelligent Information Systems*, 10(3), p.281-300.

Frawley, W. J. et al. (1991). Knowledge discovery in databases: an overview. In: Piatetsky-Shapiro, G. & Frawley, W.J., eds. *Knowledge discovery in databases*. Menlo Park: AAAI/MIT Press. 1991, p.1-30.

Furnas, G.W. et al. (1987). The vocabulary problem in human-system communication. *Communications of the ACM*, 30(11), p.964-971.

Gaizauskas, R. & Wilks, Y. (1998). Information extraction: beyond document retrieval. *Journal of Documentation*, 54(1), p.70-105.

Ingargiola, G. (1996). Building classification models: ID3 and C4.5. http://www.cis.temple.edu/~ingargiola/cis587/readings/id3-c45.html. (visited August 2000).

Lewis, D. D. (1991). Evaluating text categorization. In: *Proceedings of the Speech and Natural Language Workshop*. San Mateo: Morgan Kaufmann, p.312-318. http://www.research.att.com/~lewis (visited March 2000).

Lin, C.H. and Chen, H. (1996). An automatic indexing and neural network approach to concept retrieval and classification of multilingual (Chinese-English) documents. *IEEE Transactions on Systems, Man and* Cybernetics, 26(1), p.1-14. http://ai.bpa.arizona.edu/papers/chinese93/chinese93.html (visited November 1999).

Lin, S. H. et al. (1998). Extracting classification knowledge of Internet documents with mining term associations: a semantic approach. In: Croft, W. B. et al, eds. *Proceedings of the International ACM-SIGIR Conference on Research and Development in Information Retrieval* (SIGIR-98). New York: ACM Press, p.241-249.

Loh, S. et al. (2000). Concept-based knowledge discovery in texts extracted from the web. *ACM SIGKDD Explorations*, 2(1), p.29-39.

Sowa, J.F. (2000). *Knowledge representation: logical, philosophical, and computational foundations*. Pacific Grove: Brooks/Cole Publishing Co.

Wiener, E.D. et al. (1995). A neural network approach to topic spotting. In: 4th Annual Symposium on Document Analysis and Information Retrieval (SDAIR'95), Las Vegas. http://www.stern.nyu.edu/~aweigend/Research/Papers/TextCategorization (visited March 2000).