CLASITEX^{+}: A Tool for Knowledge Discovery from Texts

1. Introduction.

Last years are remarkable in the rapid growth of available knowledge through electronic media. Traditional data handling methods are becoming less and less capable to fulfil the demands of this information deluge. Therefore, several strategies have been proposed to do fast recovery, search and in general intelligent “analysis” of the information. All these strategies can lie on what is called Data Mining. Most of the existing work has been done on structured (i. e., numeric) databases. Nevertheless, a large portion of available information is in collection of texts written in Spanish, English, or other natural languages (histories, newspaper articles, email messages, web pages, etc.).

The problem to find interesting things in a collection of documents has been termed by Ronnen Feldman and H. Hirsh [1] as "knowledge discovery from text" and the term “text mining” has been used to refer to research in this arena. It is thus very interesting and worthwhile to develop tools to extract non-trivial information from a non-structured (i. e., textual) data base in a reasonable time.

Without any doubt, the most important works in text mining and text knowledge extraction are those developed by Ronnen Feldman´s group [1,2]; among them the KDT system analyzes and browses non-structured text collections. Each document in the collection is labeled by a set of keywords organized in a hierarchical structure.

The hierarchy of concepts is the central data structure in KDT´s architecture. The system considers that a concept is a key word. The concept hierarchy (i, e, the word hierarchy) is an acyclic directed graph where each concept is labeled by a unique name. An edge from concept A to B denotes that A is more general than B. This means that it works only with inclusion relations. The hierarchy contains only those concepts of interest to the user, and he builds it by hand.

The KDT system gives the user the possibility to browse the textual database selecting key words from the hierarchy and watching their distribution with respect to other classes or sets of key words. Each document is labeled by a set of concepts that are those appearing in its contents. In KDT these sets of concepts constitute the only information extracted from a document; each set denotes the joint occurrence of its members in the document.

KDT summarizes and analyzes the contents of the set of words labeling the documents, taking into account for this purpose the probability distribution of the daughter concepts.

A concept (node) C in the hierarchy denotes a discrete random variable whose possible values are its children. $P(C=c_i)$ denotes the distribution of the random variable C. The event $C= c_i$ is the proportion of documents annotated with c_i. $P(C=c_i)$ is the proportion of documents annotated with c_i among all documents annotated with any daughter of C. In this work the associated distribution is considered as a powerful way for browsing the data and for summarize texts and to identify interesting patterns in the data. KDT gives the possibility to the users for compare

distributions of similar keywords and view the results using tables and graphs. Finally KDT searches for irregular distributions, correlations, and associations based in conditions and thresholds supplied by the user.

FACT is other system developed by this group for knowledge discovery from texts. It discovers association-patterns of occurrence amongst keywords labeling the items in a collection of textual documents. In addition, when background knowledge is available about the keywords labeling the documents, FACT uses this information to specify constraints on the desired results of the query process. This system takes as input three sources of information. The first is a collection of textual documents on which the discovery process takes place. Each document must be labeled by a set of keywords representing the topics of the document, since this approach begins with the assumption borrowed from the Information retrieval literature. In addition to that, FACT also takes as input background knowledge about the keywords for its discovery process. To be usable such knowledge must define unary and binary predicates over the keywords labeling the documents, representing properties and relations between them. Finally, FACT allows to the user specify a query using a keyword and predicate vocabulary via a collection of menus in a simple graphical user interface.

In both before mentioned systems the document collection must be labeled by keywords. In the case of KDT a typical work session start either loading a class hierarchy from a file or by building a new hierarchy based on the collection of tags of all the documents. It is very important emphasizes that proceed in this way occasions that words equally important that the keywords and that was not considered remains out side of the process of discovery.

On the other hand FACT also requires that the collection of documents be labeled by keywords since the process of co-ocurrence discovery is realized on the basis of this sets of keywords. Something that should note is in the case of this system that does not consider the document in the discovery process. Everything turns around the keywords (more frequent words), so if the indexing process was not carry out conveniently we can not guarantee that the final result be the expected by the user.

In this work we present a system that discovers the concepts (themes or topics) most important treated by a written document in English or Spanish [4], this system works on the basis of trees of concepts. Besides it finds the relationship between the most important concepts in the text computing the co-ocurrence of apparition (of the most important concepts) in the sentences (paragraphs, sections, etc.) which conforms the texts. The system can give us a co-ocurrence distribution map of the most important concepts in the document. Finally this system in the discovery process of the most important concepts in a document read or traverses the text completely.

2. Clasitex^{+}.

CLASITEX^{+} is a system that analyzes a text in Spanish or English and discover the principal themes that are treated in the text. An important characteristic of the system is the use of a knowledge base constituted by trees of concepts.

The fundamental assumption in the system is that in a text the most repeated concept is the central theme of the same. Note that the most repeated concept is not necessarily the most

repeated word, since a concept may have associated more than one word, i.e., many words can vote by one concept.

By term we will understand a set of one or more words. So we have that a term could represent several concepts or meanings. The concepts on the other hand by definition are not ambiguous. If a term introduce polisemia and the number of different meanings is *N*, then that term generates *N* different concepts. With these concepts we will work.

A tree of concepts is an acyclic graph in which each node is a term that represent to a concept and the edges represents relationships between the concepts. Examples of the concepts considered by CLASITEX$^+$ are “home”, “Mexican Revolution”, “National Polytechnic Institute”, “Discoverer of the America”, etc.

In CLASITEX$^+$ text files are used to represent the trees of concepts. Each file represents a subtree with depth one, where the file name is the father’s name in the subtree and each one of the children appears in a line of the text file, as is showed in Fig. 1. So we have subtrees of a single level (depth one) in each file. The way in order to create new subtrees is the following:

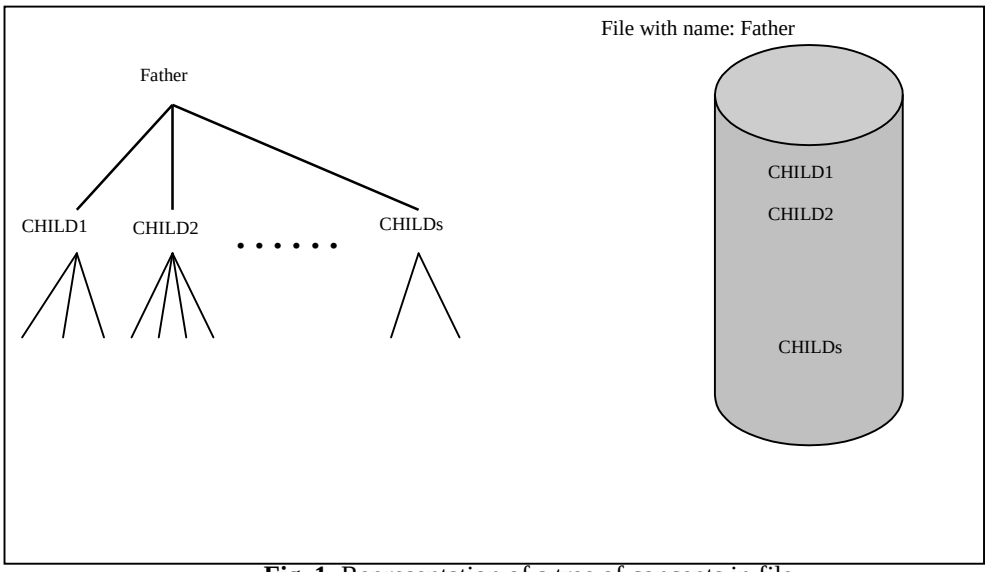

Fig. 1. Representation of a tree of concepts in file.

1. We have various directories named arbol1, arbol2, arbol3, etc., in directory arbol1 are all the trees whose children consists of a single word, in directory arbol2 are all the trees whose children have two words, and so successively.
2. If exists words of different longitude in a subtree, then in the respective directories we should create files with the same name
3. The files should be created with the following format:

<word><dot>
<word><space><word><dot>
<word><space><word><space><word><dot>, etc.

After creating the trees, are created other files named dictionaries, one for each arbol directory in dependence of the concept longitude. In these files are all the concepts in alphabetical order and the list of all the concepts that votes by him (the list of all their children). Besides, It exists terms that we know priori does not have any meaning (item, personal pronouns, prepositions, etc). These are placed in other file and do not vote for no one.

The trees were generated in this way in order to consider the polisemia problem in the terms. So, since a term may have different semantic meanings, each one of them will represent one different concept. For example the term star have meanings in the sense: astronomy, badge, famous person; then were utilized the concepts astronomy-star, badge-star, famous-person-star and were constructed the respective trees for each concept.

The considered semantic relationships in CLASITEX$^+$ to construct the trees of concepts are the following: inclusion, ownership, synonymous, conjugations, suggests, evoke, etc.

The most important module in CLASITEX$^+$ is where the analysis of the document is realized. By analysis of a document we mean to say the determination of the principal themes (concepts) in a document. In this task are considered all the trees of concepts that has been given like knowledge base for the system, and the system gets

---

Input: File with text to analyze.
DICCIONARIO-*s*= {(*c*, *concept*) / |*c*|=*s*}
where: *c* is a string of characters
|*c*| is the number of terms separated with one space in the string *c*.
DICCIONARIO-*s* is a file containing all the terms *c* such that |*c*|=*s* associated to concept.
SIN_SENTIDO is a file containing without sense terms.
Output: File (.cue) containing the concepts found in the text and their corresponding voting.
File (.res) containing the concepts for the which a term of the text votes.
File (.des) containing without sense terms in the text.
Variables: *cadena* is a string of characters.
s is |*cadena*| (number of characters in *cadena*).
índice is a pointer to the text.
N is the maximal longitude of a term.

Step 1: Pointing the variable *índice* to the beginning of the file to analyze.
Step 2: While not end of file
Step 2.1: *s* ← *N*.
Step 2.2: If *s*>0 then do
Step 2.2.1: Take *cadena* (a string) of longitude *s* starting from *índice*.
Step 2.2.2: If *cadena* exists in DICCIONARIO-s then
voting for the respective node.
índice ← *índice* + *s*.
go to step 2.1
Step 2.2.3: Else
Step 2.2.3.1: If (*s*=1) and (*cadena* exists in SIN_SENTIDO) then
índice ← *índice* + *s*.
go to step 2.1
Step 2.2.3.2: Else
write (*cadena*).
s ← *s*-1.

go to step 2.2.

Step 2.3: Else

índice ← *índice*+1

go to step 2.1

Fig. 2. Algorithm for discovers the most important concepts in a text.

the main concepts that appears in the text. For this task is necessary travel the complete document, and we should search not only isolated words but pair, trios, quartets of words, in general terms, verifying for each term if it denotes or not a concept. If denotes some concept, then one vote is given to the corresponding concept. As final result of this process we will have some concepts receive more votes than other and are precisely these which without a doubt constitutes or denotes themes in the text, the algorithm in CLASITEX$^+$ is given in Fig. 2.

The CLASITEX$^+$ system was programmed in C standard so is very portable. Besides, this system at the same time that reads or traverse a document, process it, i.e., eliminates numbers, punctuation sings and converts to lower letters the capital letters, so it is not necessary create temporal additional files.

In order to improve the speed in the voting process is convenient maintain the dictionaries of concepts and the without sense terms in memory. In this way the access to hard disk diminishes, and this benefits the velocity of the system. In order to maintain the dictionaries in memory was proposed the following data structure: An array of indexes with the letters from *a* to *z* (in ASCII code), all the combinations of two letters, the accented vowels and the letter ñ (in the Spanish case) was considered. Each one of these index points to a single link list whose nodes contain a concept that begins with these letters and besides all the concepts that vote for it, if some one index does not have associated concepts to him, this point to null. The before mentioned structure is showed in Fig. 3.

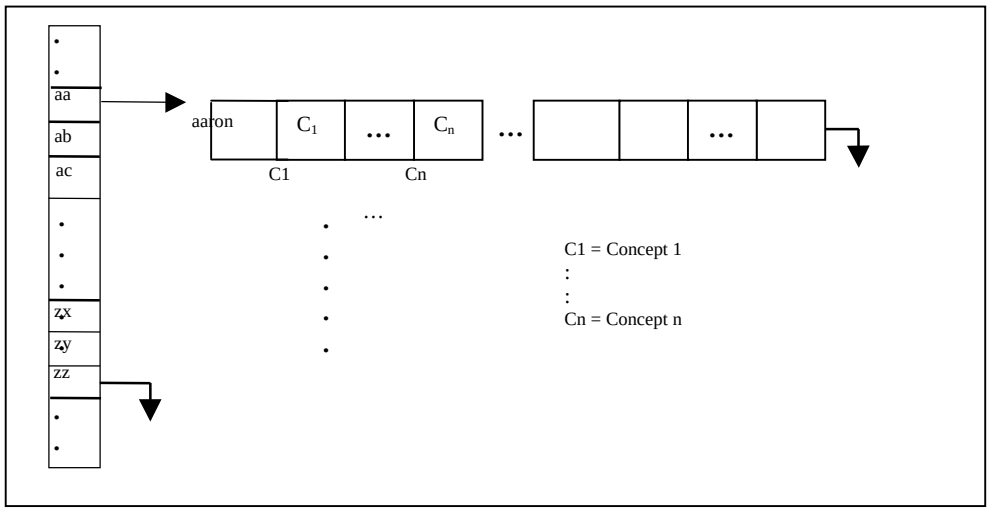

Fig. 3. Data structure for handling concepts in memory.

As result of the discovery process of the most important themes treated in a document, CLASITEX$^+$ gets a set of concepts ordered of the most frequent at least frequent. In order to discover what concepts are related, the system carries out an analysis in the co-ocurrence of apparition of the most important concepts in the sentences conforming the text. Here we are supposing if two concepts appear in the same sentence, then these are related, and if the frequency of co-ocurrence of these concepts is high then we can affirm these concepts are very related in the document. The analysis of the co-ocurrence may be carrying out by pairs, trios, and quartets of concepts in the sentences. CLASITEX$^+$ can give us a co-ocurrence distribution map of the most important concepts and their relations in the sentences, paragraphs, and sections.

The system considers the accents and the letter ñ (in Spanish case), eliminating the ambiguity that is produced in some concepts when are not considered.

The current number of concepts in Spanish that handle the system is approximately 72,000 that include areas such as: Biology, Computation, Physics, Mathematics, Medicine, Geography, etc. These concepts are of general knowledge, are not specialized.

The process to discover the most important themes in a document discussed previously is very general, so it is possible apply this same methodology independently of the idiom. If we want to find the important themes that treat a document written in English, only we need the trees of concepts in English. In CLASITEX$^+$ was considered the option for work analyzing documents in English by the system, so now the system handles 83,513 concepts of general knowledge in English.

3. Examples.

Due to space limitations we show the results of CLASITEX$^+$ in a little simple text taken of "Discover the world of science" magazine [3]. Naturally that the introduced methodology has a lot value when we have many documents and these have a large content.

Table 1. Concepts discovered by CLASITEX$^+$.

space-exploration&rocketry. 18	tools&equipment. 3
celestial-bodies. 17	tools,tests,units&scales. 3
celestial-phenomena&points. 6	large,tall,fat.. 3
electricity&electronics. 5	genetics,heredity&evolution. 2
principles-of-mechanics,waves. 4	fabrics&cloth. 2
beautiful,attractive,well-formed. 4	quantities-relationships. 2
types-of-ship&types-of-boats. 4	labor. 2
maps&cartography. 3	the-earth. 2
earth's-atmosphere. 3	materials,formations. 2
publishing. 3	elements. 2
forecasting&meteorology. 3	measures&standards-of-time. 2

SATURN, 2004

Ten or twenty years a go, interplanetary space probes where built like battleships: big, rugged, bristling with instruments - and costing a boatload of money. Although NASA has been phasing out such missions, only in October did it finally launch its last: the Cassini probe to Saturn.

By 2004, if all goes well, Cassini will park itself in orbit around Saturn loop around and around, taking readings and snapping close-ups of the planet, its gossamer rings, and its 18 moons - the some sort out work its cousin Galileo is now doing at Jupiter. Like Galileo, Cassini is two probes in one. While the main craft orbits Saturn a second probe come the European - built Huygens, will detach and fall into the atmosphere of Titan, Saturn's largest moon. Titan is a world into itself, nearly as big as Mars, it has an atmosphere that astronomers think is laced with organic chemicals - the building blocks of life. "There are only a few solid bodies in the solar system with thick atmospheres - Earth,Venus, and Titan", says planetary scientist Jonathan Lunine the University Arizona in Tucson. "A Titan is the best model for the Earth prior to the time when life began".

Cassini also resembles Galileo in that it carries radioactive plutonium - 72 pounds of the poisonous stuff, which provides heart to power the probe beyond Mars. The risk of that much plutonium being accidentally released into the atmosphere, either at launch or when Cassini flies by Earth for a gravity assist in August of 1999, drew a great deal of protest, as did the 1989 launch of Galileo. But in a perfect reply, Cassini headed of into deep space without a serious hitch.

The most important concepts discovered by CLASITEX$^+$ are the showed in the Table 1.

CLASITEX$^+$ takes 1.15 seconds in order to discover the most important concepts. From the previous result we have that the most frequent concepts are: 1)space-exploration&rocketry, 2)celestial-bodies, 3)celestial phenomena&points and 4)electricity&electronics.

Now computing the co-ocurrence between pairs of concepts we can find how are related these concepts.

Table 2. Co-ocurrence of apparition of the most important concepts in the sentences conforming the text.

Sentences	Pairs of concepts				
	1,2	1,3	1,4	2,3	3,4
Ten or twenty years a go…					
Although NASA has been…	1	1			
By 2004, if all…	1	1		1	
Like Galileo…					
While the main…	1	1	1	1	1
Titan is a world…	1			1	
"There are only….	1			1	
"A Titan is…					
Cassini also…					
The risk of…	1	1		1	
But in a perfect…					
Total	6	4	1	5	1

So from the analysis of co-ocurrence (see Table 2), we can see that the related concepts are:

space-exploration&rocketry and *celestial-bodies*
celestial-bodies and *celestial-phenomena&points*
space-exploration&rocketry and *celestial-phenomena&points*

So the final user can easily see that the analyzed text treat about the space exploration of celestial bodies, also about celestial bodies, phenomena and celestial points, and finally about space exploration of phenomenon and celestial points. All the before in automatic way and in little seconds. Besides CLASITEX$^+$ gives the possibility of locating exactly within the text, the zones where the most important related concepts appear.

4. Conclusions.

In this work was presented the CLASITEX$^+$ system, where some strategies were introduced for handle trees of concepts in memory and increase the speed in the analysis of a document (written in Spanish or English). The system discovers the most important concepts (main themes) treated in a document, and also finds which of those concepts are related computing the co-ocurrence of

apparition of the same in the sentences of the document, in an acceptable time. Other important aspect in the system is the amount of concepts in Spanish and English.

This work constitutes a step more in the analysis of texts or non-structured data. In future work improvements in the strategy to voting will be introduced, we will introduce a semantic similarity measure between concepts that consider the relative position of the concepts within the tree. This will help us to give a weight to each concept in the voting process, this weight will be in function of the semantic similarity and, therefore, will reduce the ambiguity problem. Besides will be solved problems such as: establish order between the secondary themes of a document; prepare a summary of the fundamental contents in a document; describe tendencies or the conceptual evolution in time of an information source emitting a set of documents; reveal the conceptual nexus between them, etc.

This work was partially financed by Dirección de Estudios de Posgrado e Investigación del Instituto Politécnico Nacional and the CONACyT Projects No.3757P-A9608 and REDI of Mexico.

References

[1]Feldman R. and Hirsh H. “Mining association in text in the presence of background knowledge”. In the Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, Portland Oregon, August, 343-346 (1996).

[2]Feldman R. and Ido Dagan. “Knowledge Discovery in textual databeses (KDT) ”. In the Proceedings of the first Int. Conf. On Data Mining and Knowledge Discovery (KDD95), pp. 112-117, Montreal, Aug 1995.

[3]Flamsteed S. “Saturn, 2004”. Discover, The world of science, pp. 76, January 1998.

[4]Guzmán A.A. “Finding the main themes in a spanish document”. Journal Expert Systems and Applicacions, Vol.14, No.1, January (1998).

[5]Rijsbergen C.J., et al. “Information retrieval”. Second Edition 1979, http://www.dcs.gla.ac.uk/Keith/Preface.html.

6]Salton G. “Automatic Text Processing: The transformation, Analysis and Retrieval of Information by Computer”. Addison-Wesley Publishing Company (1989).