Computación y Sistemas

Document Indexing with a Concept Hierarchy

Índice de documentos con una jerarquía de conceptos

A. Gelbukh, G. Sidorov, and A. Guzmán-Arenas

Natural Language Processing Laboratory,
Center for Computing Research (CIC), National Polytechnic Institute (IPN),
Av. Juan de Dios Batiz s/n esq. Mendizabal, Col. Zacatenco, CP 07738, DF, Mexico.
e-mail: gelbukh*gelbukh.com, sidorov*cic.ipn.mx, a.guzman*acm.org;
www.Gelbukh.com

Artículo recibido en agosto 25, 2002; aceptado en febrero 20, 2005

Abstract

Given a large hierarchical concept dictionary (thesaurus, or ontology), the task of selection of the concepts that describe the contents of a given document is considered. A statistical method of document indexing driven by such a dictionary is proposed. The method is insensible to inaccuracies in the dictionary, which allow for semi-automatic translation of the hierarchy into different languages. The problem of handling non-terminal and especially top-level nodes in the hierarchy is discussed. Common sense-complaint methods of automatically assigning the weights to the nodes and links in the hierarchy are presented. The application of the method in the Classifier system is discussed.

Keywords: document characterization, document comparison, ontology, statistical methods.

Resumen

Se considera la tarea de la selección de los conceptos que describen el contenido de un documento dado. Los conceptos se eligen de un diccionario jerárquico grande (un tesauro, o bien una ontología). Se propone un método estadístico para crear un índice de los documentos, guiado por tal diccionario. El método es robusto en cuanto a los errores en el diccionario, lo que permite traducir tal diccionario semiautomáticamente en varios lenguajes. Se discuta el problema del uso de los nodos no terminales y especialmente de los nodos de alto nivel en la jerarquía. Se presentan los métodos para ponderación automática de los nodos y vínculos en la jerarquía de la manera que coincide con los criterios del sentido común. Se discuta la aplicación del método en el sistema *Classifier*.

Palabras clave: caracterización de documentos, comparación de documentos, ontología, métodos estadísticos.

1 Introduction

We consider the task of indexing a document with concepts as mapping the document into the concept dictionary, assigning to each concept in the dictionary a value that reflects its relevance for the given document. Thus, the document is represented by a histogram of its topics. Say, a newspaper article can be about *industry* (60%), *transport* (20%), *science* (10%), etc. Note that these are concepts included in the dictionary rather than the key words directly mentioned in the document; what is more, the document might not contain the word *transport* at all, but instead contain the words *trains*, *railways*, etc.

Such a representation of documents is important for information retrieval (Chakrabarti *et al*. 1997), document classification, text mining (Feldman and Dagan 1995), investigation of document collections (Light 1997), text understanding, etc.

In document retrieval, the documents are scored by the correspondence of their main topics to the user's request. In text mining, data mining techniques are applied to discovering trends and deviations of the topics of discussion in the newspapers. In text understanding, topic detection allows selecting the language model (Seymore and Rosenfeld 1997). In document classification and text segmentation (Ponte and Croft 1997), topic detection has been the object of extensive research in recent years.

A large core of research has been devoted to automatically learning the classification rules using statistical and linguistic methods (Apté *et al*. 1994; Bharat and Henzinger 1998; Cohen and Singer 1996), machine learning methods (Koller and Sahami 1997), neural networks, self-organizing maps (Hyötyniemi 1996) and connectionist models (Le *et al*. 1994). In the majority of these studies, the task of automatic construction of the topic hierarchy is considered. In this article,

Alexander Gelbukh, Grigori Sidorov, Adolfo Guzman-Arenas. Document Indexing with a Concept Hierarchy.
Computación y Sistemas, Ibero-American Computing Journal, 2005, to appear.

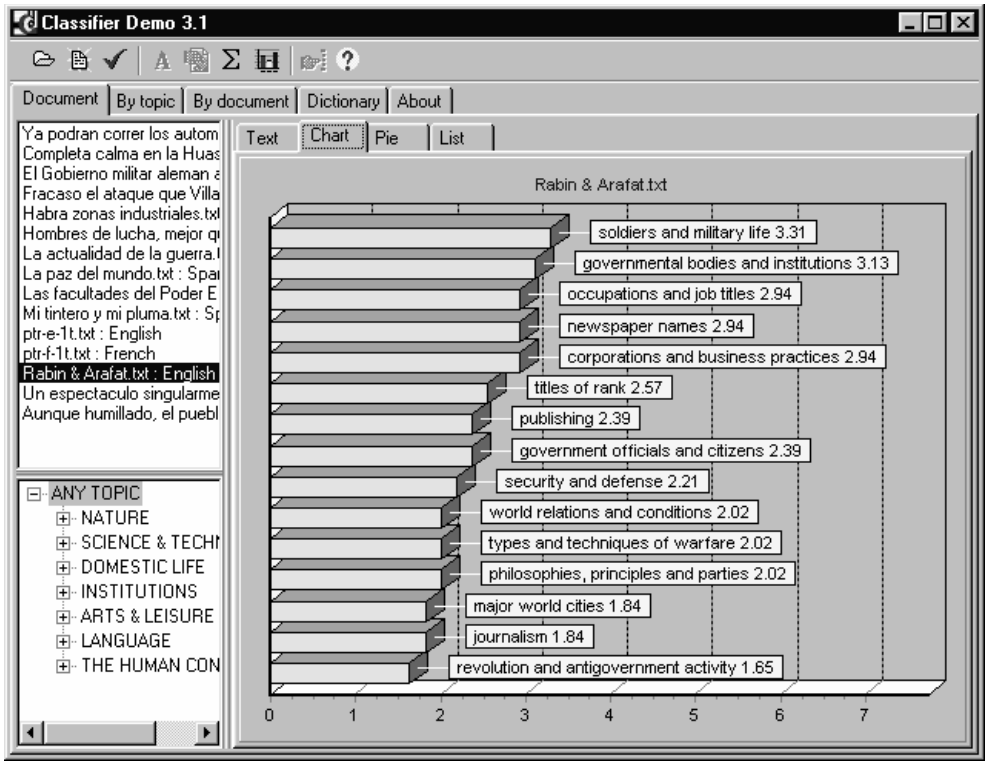

Figure 1. Topic histogram for a document.

however, we consider the (non-weighted) topic hierarchy to be pre-defined, and concentrate on its using and on assignment of the weights to the nodes and links.

The problems arising in the compilation and use of a concept hierarchy depend dramatically on its size. In some applications, there is a small set of predefined topics, and a typical document is related to only one topic. For example, this is the case for a governmental reception office where the complaints it receives from the citizens are to be classified to send them to exactly one of the departments of *police*, or *health*, or *environment*, etc.

However, in the case of open texts, such as Internet documents or newspaper articles, the set of possible topics is large and not so well defined, and the majority of the documents are related to several or many topics at the same time. This leads to the necessity of some structuring of the set of topics. The most natural structure for the concepts is a hierarchy. For example, if a document is related to the narrow topics *elections*, *government*, and *party*, then it can be classified as a document on *politics*.

Thus, though most of existing dictionary-based systems use "flat" topic dictionaries – keyword groups without any hierarchical structure – in this paper we use a hierarchical dictionary and specifically address the issue of determining the contribution of the top-level concepts. Such a problem does not exist in the "flat" document categorization dictionaries.

We consider the list of topics to be large but fixed, i.e., pre-defined. Our indexing algorithm does not obtain the topics directly from the document body; instead, it relates the document with one of the topics listed in the system dictionary. The result is, thus, the measure (say, in percents) of the corresponding of the document to each of the available topics. Unlike the traditional categorization approach, we consider "fuzzy" categorization, when a document can be indexed with many categories with their corresponding weights. In Figure 1, a screen shot of our program, CLASSIFIER, is shown with a histogram of the topics of a Spanish document.

A problem arises of the optimal, or reasonable, degree of detail for such categorization. For example, when describing the Internet news for an "average" reader, the categories like *animals* or *industry* are quite appropriate, while for the description of articles on zoology such a dictionary would give a trivial answer that all documents are about *animals*. On the other hand, for "average" reader of Internet news it would not be appropriate to categorize the documents by the topics *mammals*, *herptiles*, *crustaceans*, etc., since such a description is too detailed for such a user.

In this paper, we will discuss the structure of the topic dictionary, the choice and use of topic weights, and some practical aspects of compilation of the hierarchical concept dictionary.

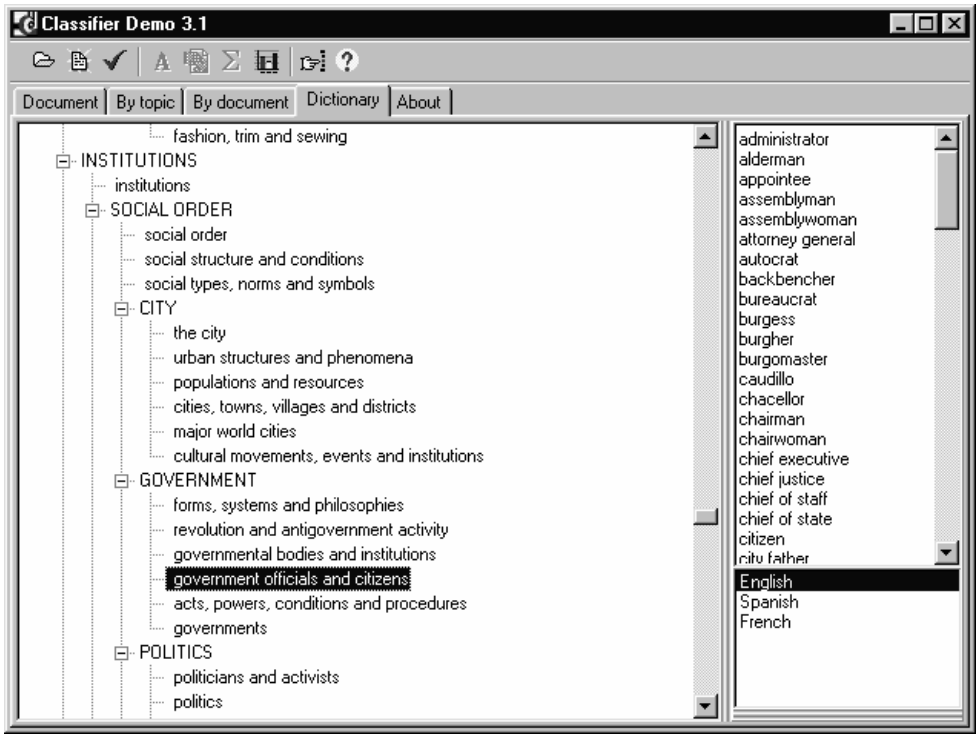

Figure 2. Hierarchical dictionary used by the system.

2 Concept Hierarchy

In (Guzmán-Arenas 1997 and 1998), it was proposed to use a hierarchical dictionary for determining the main themes of a document. The proposed algorithm is based on this idea. Unlike some other methods of indexing (Niwa et al 1997), our algorithm does not obtain the candidate topics directly from the body of the document being analyzed. Instead, it relies on a large pre-existing dictionary of topics organized in a tree. Non-terminal nodes of this tree represent major topics, such as *politics* or *nature*. The terminal nodes represent the narrowest topics such as *elections* or *crocodiles*.

Terminal topics are associated with so-called keyword groups. A keyword group is a list of words or expressions related to the situation described by the name of the topic. Such words and expressions are directly used in the text. For example, the topic *religion* can be associated with the words like *church*, *priest*, *candle*, *Bible*, *pray*, *pilgrim*, etc.
Note that these words are connected neither with the headword *religion* nor with each other by any "standard" semantic relation such as subtype, part, actant, etc. This makes compilation of such a dictionary easier than that of a real semantic network dictionary. However, such a dictionary is not a "plain" variant of a semantic network such as WordNet, since some words are grouped together that have no immediate semantic relationship. Thus, such a dictionary cannot be obtained from a semantic network by a trivial transformation.

Figure 2 shows another example of a dictionary entry. Technically, our CLASSIFIER program manages contact word combinations in the same way as single words.

Though the concepts are organized in a tree, a keyword can belong to several concepts. This can be due to either homonymy of the word (Krowetz 1997): e.g., *bill* belongs to *money*, *law*, *tools*, *birds*, or due to intersection of the topics: e.g., *girl* belongs to *children* and *women*.

3 The Naïve Algorithm: No Weights

The algorithm of document indexing with the concept thesaurus consists of two parts: individual (leaf) topic detection and propagation of the topics up the tree.

###

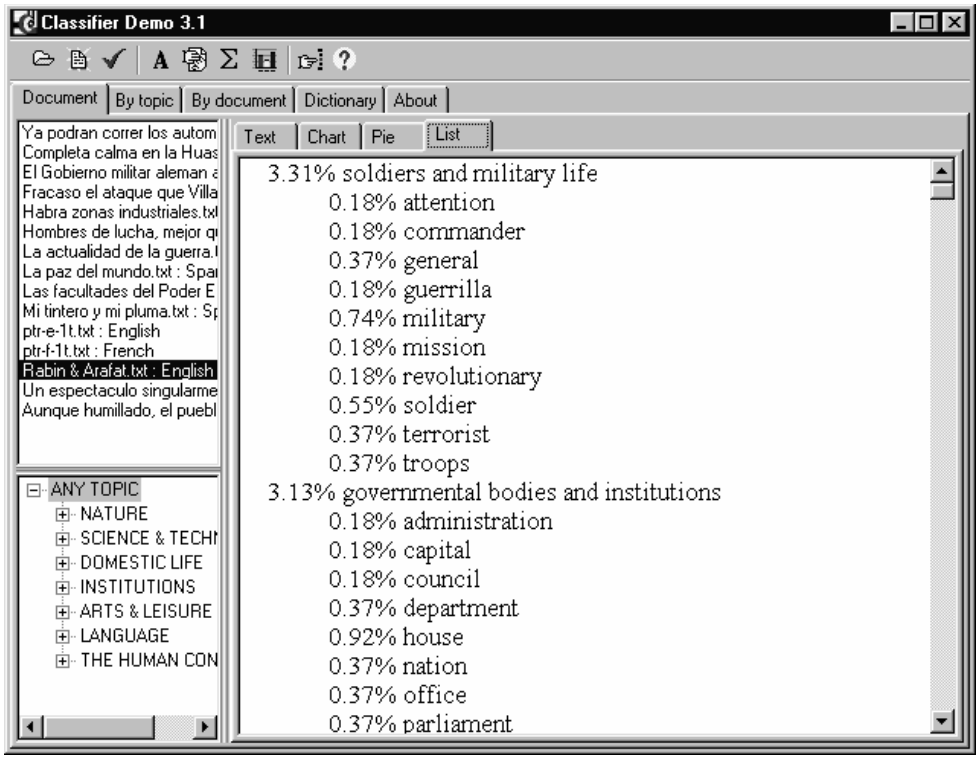

Figure 3. Counting keywords in a document.

The first part of the algorithm is responsible for detection terminal topics, i.e., for answering, individually for each terminal topic, the following question: To what degree this document corresponds to the given topic? In our current implementation, this is done basing on a plain list of words corresponding to the topic, see **Figure 3**. However, in general, a topic can be associated with a procedure. For example, to detect that a document represents an application form relevant to some department of a government office, it may be necessary to analyze the format of the document.

In our implementation, for each keyword group, the number of occurrences of the words corresponding to each (terminal) topic is determined. These numbers are normalized within the document, i.e., divided by the number of words in the document. The accumulated number of occurrences is considered the measure of the correspondence between the document and the topic. Note that the values for this measure of relevance are not normalized since the topics are not mutually exclusive.

The second part of the algorithm is responsible for the propagation of the found frequencies up the tree. With this, we can determine that, say, a document mentioning the terminal topics *mammals*, *herptiles*, *crustaceans*, is relevant for the non-terminal topic *animals*, and also *living things*, and also *nature*.

The difference between indexing with a plain list of terminal topics and with propagation of the frequencies to non-terminal topics can be seen in **Figure 1** and Figure 4. Actually, in **Figure 1**, only terminal concepts are shown, while Figure 4 shows all concepts, including non-terminal ones, for the same document. Though the main terminal topic for this document is *soldiers and military life*, the main topic in the whole tree is *INSTITUTIONS*.[1]

Propagation of the frequencies is crucial for the whole idea of our method. It is necessary to make use of the non-terminal nodes of the hierarchy and to generalize the contents of the document to a degree allowing for its matching with the user's queries containing more general words than the ones mentioned directly in the document. However, it presents the problem of overgeneralization: applied in the naïve way described here, it always assigns the greatest relevance to the top-level concepts, so that any document is indexed with the concepts *object*, *action*, etc., as its main topics.

As an example of this problem, we can see that in Figure 4, the concept *ANY TOPIC* has unreasonably high ranking. Below we will show how to cope with this problem.

[1] In Figure 4, the non-terminal topics are shown in capital letters.

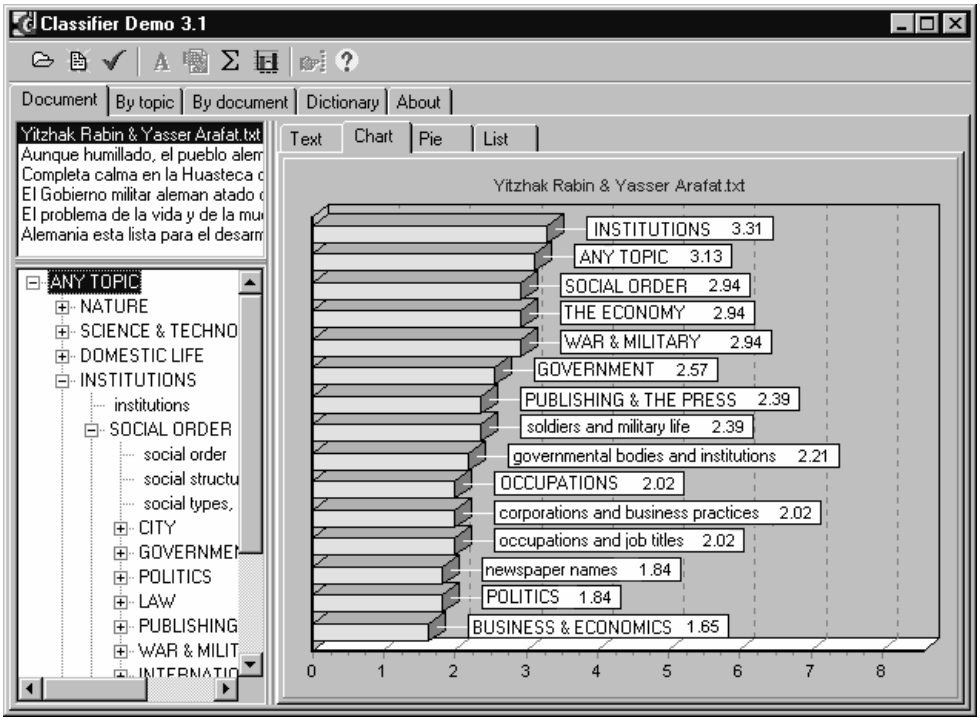

Figure 4. Non-terminal concepts in the index.

4 Relevance and discrimination weights

In the algorithm described above, simple word lists were used to count the frequencies reflecting the relevance of the topic for a document. Thus, a word in the document could be counted either as one occurrence of the corresponding topic or zero, which is too rigid.

To define quantitative measures of relevance of the words for the topics and quantitative measure of importance of the nodes of the hierarchy, some numeric weights can be used by the algorithm. There are two kinds of such weights: the degree of the connection between a keyword (or a subtopic) and the corresponding topic is associated with a link in the hierarchy, while the measure of importance of an individual concept for the user is associated with the individual node.

The classification algorithm takes into account these weights. Namely, for the accumulated relevance of the topics, it multiplies the number of occurrences of a word (or subtopic) by the weight w_k^j of the link between the word and the topic, and then multiplies the result by the weight w^j of the topic itself.

4.1 Relevance and discrimination weights

The first type of the weights is associated with the links between words and topics or between the nodes in the tree.[2] For example, if the document mentions the word *carburetor*, is it about *cars*? And if it mentions the word *wheel*? Intuitively, the contribution of the word *carburetor* into the topic *cars* is greater than that of the word *wheel*: if the document mentions *carburetor*, then it's almost surely is about cars, but if it mentions *wheel*, then it is possibly about cars but could be about *clocks*, *songs*, *pottery*, etc. Thus, the link between *wheel* and *cars* is assigned a less weight than for *carburetor* and *cars*. The algorithm of classification takes into account these weights when compiling the accumulated relevance (frequency) of the topics.

[2] Actually, the former type is a kind of the latter since the individual words can be considered as terminal tree nodes related to the corresponding topic.

It can be shown that the weight w_k^j of such a link (between a keyword *k* and a topic *j* or between a topic *k* and its parent topic *j* in the tree) can defined as the mean relevance of the documents containing this word for the given topic:

$$w_k^j = \frac{\sum_{i \in D} r_i^j n_i^k}{\sum_{i \in D} n_i^k}, \tag{1}$$

by all the available documents *D*, where r_i^j is the measure of relevance of the document *i* to the topic *j*, and n_i^k is the number of occurrences of the word or topic *k* in the document *i*.

This equation can be used for automatic training of the dictionary given a document collection marked up with the relevance of the topics. Unfortunately, we are not aware of any reliable algorithm of automatic detection of the measure of the relevance of r_i^j in an independent way. Thus, to apply this equation directly or training the dictionary, such a measure r_i^j is to be estimated manually by the expert.

However, in practice such a work is usually too expensive. As a practical alternative, it is often possible for the expert to estimate the weights w_k^j intuitively at the stage of preparation of the dictionary. The choice of the weight is based on the frequency of appearance of the word in "general" documents from the control corpus of the texts on "any" topic; in our case such texts were the newspaper issues.

As another practical approximation, for narrow enough themes we can assume the hypothesis that the texts on this topic never occur in the control corpus (newspaper mixture). Then, given the fact that we have included the word in the dictionary and thus there is at least one document relevant for the given topic, we can simplify the expression for the weights as follows:

$$w_k^j = \frac{1}{\sum_{i \in D} n_i^k}, \tag{2}$$

since the numerator of the quotient in (1) in case of narrow topics can be considered to be 1. Not surprisingly, this gives the weight of the word related to a specific topic to be the less the more its frequency; for example, the articles *a* and *the* have a (nearly) zero weight for any topic, while the word *carburetor* has a high weight in any topic in which it is included.

Sometimes a rare enough word, say, a noun *bill*, in its different senses is related to different topics (*money*, *law*, *birds*, *geography*, *tools*). For a more accurate analysis, some kind of competition between senses of the word for a specific occurrence in the document is introduced. For this, the senses of the word are marked in the topic dictionaries (as $bill_1$, $bill_2$, etc.), and the weights of occurrences of such a word are to be normalized by its different senses (though the occurrences of the same sense are independent in different topics), with the weigh of an individual sense in each document being proportional to the relevance of the document for the given topic:

$$\begin{aligned} w_k &\sim \sum_j r_i^j w_k^j \\ \sum_k w_k &= 1 \end{aligned}, \tag{3}$$

where w_k is the weight of the *k*-th sense of the given occurrence of the word in the given document *i*, w_k^j is the weight of the link between this sense of the word and the topic *j*, the summation in the first equation is made by all the topics, and in the second by the senses of the given word.[3] Since r_i^j in its turn depends on w_k, to avoid iterative procedure, in practice we calculate r_i^j based on equal weights w_k.

[3] However, this technique is not very important for most cases, since usually it does not change the order of the topics for a document, but only makes the difference between different topics more significant. Since in many cases it requires manual marking up the senses (for the words for which such information is absent in the dictionary), we use it rarely in our program; for the majority of the words we do not distinguish senses.

Note that the weights of the links are a natural part of the concept hierarchy itself. This component imparts a quantitative character to the hierarchy. However, here we discuss the issue of calculating the weights because of two reasons. First, there are available concept hierarchies, such as Roget thesaurus, WordNet (1998), Factotum SemNet (Cassidy 2000), etc. However, these dictionaries do not include any quantitative information.

Second, the relevance weights, generally speaking, depend on the training set – the collection of general texts. For example, for application of our indexing method to a collection of technical articles, the link between *wheel* and *song* would have much less weight than for its application to the general Internet documents. This effect will be described in more detail in the next section.

4.2 Discrimination Weights

The classification algorithm described above is good for answering the question "is this document about *animals*?" but not the question "what about is this document?". Really, as we have mentioned, with such an approach taken literally, the answer will be "all the documents are about *objects* and *actions*," the top nodes of the hierarchy. However, a "reasonable" answer is usually that a document is about *crustaceans*, or *animals*, or *living things*, or *nature*, depending on the user's needs and level, i.e., on the degree of details to which the user is interested in the area.

Thus, we suggest that the answer to the question "what about is this document?" depends on the user. For example, if the document mentions *lobsters*, *shrimps*, *crabs*, and *barnacles*, then for a biologist the answer *crustaceans* would be the best, while for a schoolchild the answer *biology* is better, and for an average newspaper reader, the answer *nature*.

How can we guess this without having to explicitly ask the reader? Asking the reader about the desired detail degree is not a solution because, first, he or she will probably even not understand the question, and, second, it is not possible for the reader to quantitatively specify the importance of hundreds of topics in the hierarchy. Thus, an automatic way of assigning the importance weights is necessary.

Our hypothesis is that the "universe" of the reader is the base of the documents to which he or she applies the search or classification. In other words, we assume that the reader is a specialist in the contents of the current database being indexed. Thus, the weights of the relevance of topics in our system depend on the current database.

The main requirement to these weights is their discrimination power: an important topic should correspond to a (considerable) subset of documents, while the topics that correspond to nearly all documents in the data base are probably useless, as well as too narrow topics that correspond to few documents in the base. Thus, the weight w^j of a tree node j can be estimated as the variation of the relevance r_i^j the topic over the documents of the database. A simple way to calculate such a discrimination power is to simply measure it as the dispersion:

$$w^j = \sum_{i \in D} \left(r_i^j - M \right)^2$$
$$M = \frac{\sum_{i \in D} r_i^j}{|D|} \tag{4}$$

where M is the average value of r_i^j over the current database D, and r_i^j is determined by the former algorithm, without taking into account the value of w^j.[4]

With this approach, for, say, a biological database, the weight of the topics like *animals*, *living things*, *and nature* is low because all the documents equally mention these topics. On the other hand, for newspaper mixture their weight is high.

5 Applications

With the approach described above, we have implemented in the system CLASSIFIER several useful functions.

The system can index the document with its main topics, with or without the propagation of the frequencies to the non-terminal nodes, see Figure 1 and Figure 4. The system allows viewing the documents by topics, answering the question:

[4] In a more precise manner, the information theory can be applied to the calculation of the weights; we will not discuss here this idea. Yet another approach to the propagation of the frequencies to the non-terminal nodes is discussed in (Gelbukh *et al*. 1999b).

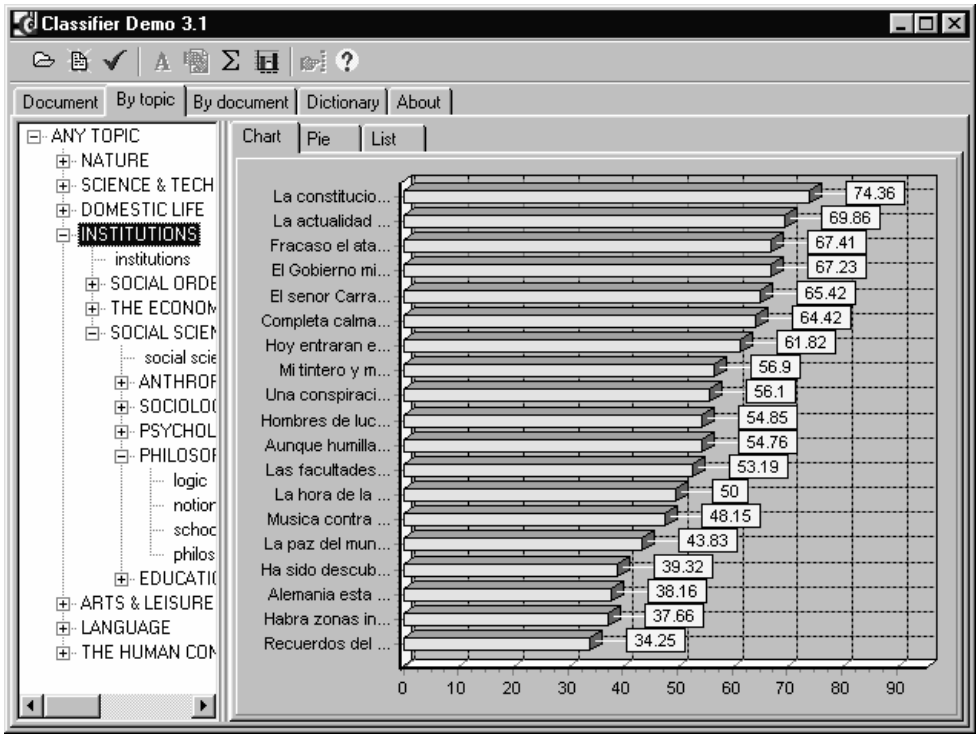

Figure 5. Documents ordered by relevance for a specific topic.

for a selected topic, what documents are the most relevant? This corresponds to the task of information retrieval, see Figure 5.

An interesting application of the method is classification of the documents by similarity to a given document, see Figure 6. The comparison is made based on the relevance of concepts from the dictionary for the two documents, i.e., on their representation with the topic histograms, see, for example, (Gelbukh *et al*. 1999a).

The hierarchical structure of the dictionary allows for taking into account the user's interests in two ways: automatically and manually. The automatic way consists in the use of the weights of the importance described in the previous section: the documents devoted to the same important topics are considered close, even if they differ in the details.

Alternatively, by manually restricting the comparison to some subtree of the concept hierarchy, the user can compare the documents with respect to a given topic. Let us consider two documents: one mentioning the use of *dogs* for *sabotage acts* and another one mentioning the *feeding* of *dogs*. For an average user, they are moderately similar, since both mention *dogs* among other things. However, from the point of view of a zoologist, they are very similar since both concern with the *dogs* (and not *cows* or *crocodiles*). For a military man, though, they are not similar at all: one of them mentions *sabotage acts*, and another one does not mention anything important. In the latter two cases, the users can choose for the comparison the sub-trees of the topics *biology* and *military and war*, correspondingly.

6 Practical Issues

The most difficult problem in application of our method is compilation of the dictionary. In our case, we compiled the basic part of the English hierarchical dictionary from various sources, including Roget thesaurus, WordNet, FACTOTUM SemNet and some other dictionaries.

However, for our goals we needed a Spanish and a French dictionary. As it turned out, automatic translation of the English dictionary gave very good results. Even a simple translation procedure that translated the words out of context (and thus did not discriminate meanings), provided us with a usable dictionaries. Further improvements to the Spanish and French dictionaries are achieved by using a context-sensitive procedure described in (Gelbukh 1999).

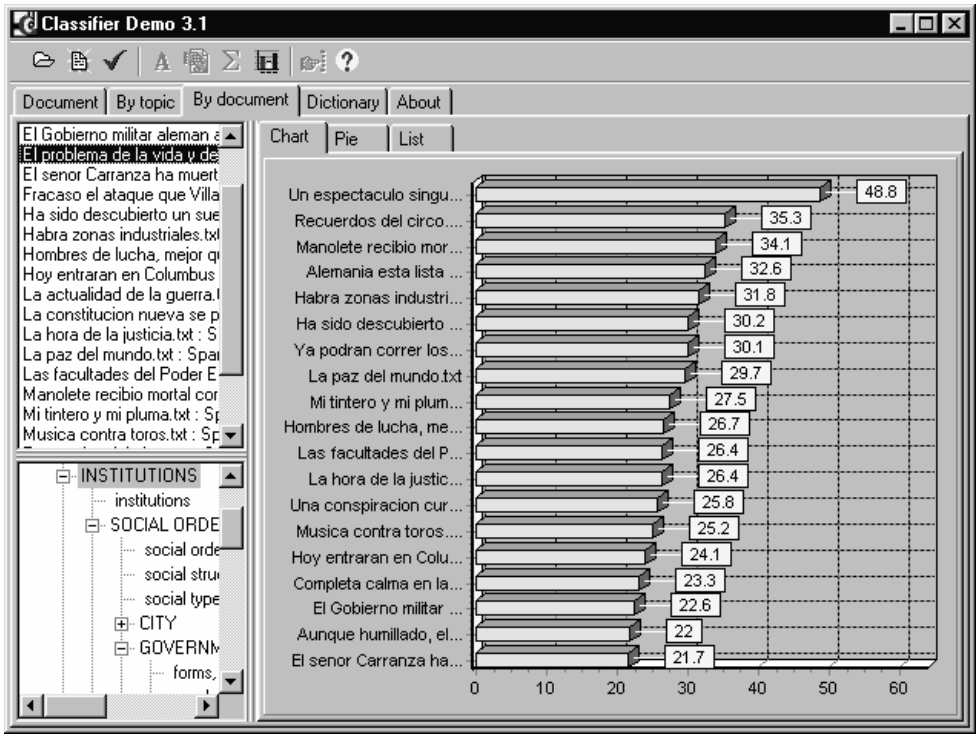

Figure 6. Similarity of the documents to a given document with respect to a given aspect (topic *Institutions*).

For the languages other than English, a powerful morphological engine has to be used to match the words in all their morphological forms to the words in the documents. However, with a simpler approach, only nouns and adjectives can be used in the dictionary, that greatly reduces the complexity of morphological processing for the languages that do not have grammatical cases. In our implementation, we kept both forms of English, French, and Spanish nouns, as well as four forms of French and Spanish adjectives, directly in the dictionary.

7 Discussion and Future Work

Generally, the results obtained in our experiments show good accordance with the classification made by human experts. However, we encountered some problems with using our method. Most of them are related with ambiguity.

Sometimes, a frequent keyword (taken out of context) proves to be important for a specific topic: the noun *well* is an important term in *petroleum extraction*, the noun *do* is a term in *hairstyles*, the noun *in* in *politics*, etc. However, the expression (1) assigns too little weight to such keywords. To solve this problem, we plan to add a part of speech tagger to our system. For a more detailed analysis, we will have to add a syntactic parser to the program; however, this would greatly slow down the system.

Obviously, this does not solve all the problems of ambiguity. As we have discussed, for the words like *bill* a sophisticated and not always reliable algorithm is used; we plan to resolve the ambiguity of this type with more intelligent methods described, for example, in (Gelbukh 1997).

Another important issue that can improve the quality of classification is anaphora resolution. With anaphoric links at least partially resolved, the pronouns can be counted as occurrences of the corresponding nouns.

8 Conclusions

We discussed a method of document indexing driven by a hierarchical concept dictionary. The method is statistical-based and involves the quantitative measure of the strength of the links in the hierarchy, as well as the weights of importance of the nodes of the hierarchy for the user. We have suggested that the latter weights depend on the database to which the indexing algorithm is applied. We have discussed the automatic procedure of assigning the corresponding weights to the links and the nodes in the concept hierarchy. The discussed methods have been implemented in a system *Classifier* for document retrieval and investigation of document collections.

Though there are some problems with the accuracy of the algorithm, the results of experiments show good accordance with the opinion of human experts. The method is practical in the sense of insensibility to inaccuracies in the dictionary and in the sense of using a dictionary with a simple structure.

The directions of further improvements to the method are related with application of intelligent linguistic methods of lexical and anaphoric disambiguation.

Acknowledgments

The work was partially supported by Mexican Government (SNI, CONACyT, CGPI-IPN).

References

Apté Ch; **F. Damerau**, and **Sh. M. Weiss**, "Automated learning of decision rules for text categorization". *ACM Transactions on Information Systems*. Vol. 12, No. 3 (July 1994), pp. 233-251.

Bharat K. and **M. Henzinger**, "Improved algorithms for topic distillation in hyper-linked environments", *21st International ACM SIGIR Conference*, 1998.

Cassidy P., "An Investigation of the Semantic Relations in the Roget's Thesaurus: Preliminary results", In: *Proc. CICLing-2000, International Conference on Intelligent Text Processing and Computational Linguistics*, IPN, Mexico, 2000, 181–204.

Chakrabarti S.; **B. Dom, R. Agrawal**, and **P. Raghavan** "Using taxonomy, discriminants, and signatures for navigating in text databases", *23rd VLDB Conference*, Athenas, Greece, 1997.

Cohen W. and **Y. Singer**, "Context-sensitive Learning Methods for Text Categorization", *Proc. of SIGIR'96*, 1996.

Feldman R. and **I. Dagan**, "*Knowledge Discovery in Textual Databases*", Knowledge Discovery and Data Mining, Montreal, Canada, 1995.

Gelbukh A., "Using a semantic network for lexical and syntactic disambiguation", *Proc. of Simposium Internacional de Computación: Nuevas Aplicaciones e Innovaciones Tecnológicas en Computación*, November 1997, Mexico.

Gelbukh A., "Syntactic disambiguation with weighted extended subcategorization frames". *Proc. PACLING-99, Pacific Association for Computational Linguistics*, Canada, pp. 244–249.

Gelbukh A., **G. Sidorov**, and **A. Guzmán-Arenas**, "Document comparison with a weighted topic hierarchy", *Proc. 1st International Workshop on Document Analysis and Understanding for Document Databases (DAUDD'99), 10th International Conference and Workshop on Database and Expert Systems Applications (DEXA)*, Florence, Italy, September 1, 1999. IEEE Computer Society Press, pp. 566-570.

Gelbukh A., **G. Sidorov**, and **A. Guzmán-Arenas**, "A Method of Describing Document Contents through Topic Selection". *Proc. of SPIRE'99, International Symposium on String Processing and Information Retrieval*, Cancun, Mexico, September 22 – 24. IEEE Computer Society Press, 1999, pp. 73-80.

Guzmán-Arenas A., "Finding the main themes in a Spanish document", *Expert Systems with Applications*, Vol. 14, No. 1/2, Jan/Feb 1998, pp. 139-148.

Guzmán-Arenas A., "Hallando los temas principales en un artículo en español," *Soluciones Avanzadas*. 1997, Vol. 5, , No. 45, p. 58, No. 49, p. 66.

Hyötyniemi H., "Text Document Classification with Self-Organizing Maps", in *STeP'96, Genes, Nets and Symbols*, Alander, J.; Honkela, T.; Jakobsson, M. (eds.), Finnish Artificial Intelligence Society, 1996, pp. 64-72.

Koller D. and **M. Sahami**, "Hierarchically classifying documents using very few words", *International Conference on Machine Learning*, 1997, pp. 170-178.

Krowetz B. "Homonymy and Polysemy in Information Retrieval", *35th Annual Meeting of the Association for Computational Linguistics*, 1997, pp. 72-79

Le D.X., **G. Thoma** and **H. Weschler**, "Document Classification using Connectionist Models", *IEEE International Conference on Neural Networks*, Orlando, FL, June 28 – July 2, 1994, Vol. 5, pp. 3009-3014.

Light J., "A distributed, graphical, topic-oriented document search system" *CIKM '97, Proceedings of the sixth international conference on Information and knowledge management*, 1997, pp. 285-292.

Niwa Y., Sh. Nishioka, M. Iwayama, A. Takano, and **Y. Nitta**, "Topic Graph Generation for Query Navigation: Use of Frequency Classes for Topic Extraction", *NLPRS'97, Natural Language Processing Pacific Rim Symposium '97*, Phuket, Thailand, Dec. 1997, pp. 95-100.

Ponte J. M. and **W. B. Croft**, "Text Segmentation by Topic", *First European Conference on Research and Advanced Technology for Digital Libraries*, 1997, pp. 113-125.

Seymore K. and **R. Rosenfeld**, "Using story topics for language model adaptation", *Proc. of Eurospeech'97*, 1997.

WORDNET, *Coling-ACL'98 Workshop: Usage of WordNet in Natural Language Processing Systems*. August 16, 1998, Université de Montréal, Montréal, Canada.

Alexander Gelbukh *was born in Moscow, Russia, in 1962. He obtained his Master degree in Mathematics in 1990 from the department of Mechanics and Mathematics of the "Lomonosov" Moscow State University, Russia, and his Ph.D. degree in Computer Science in 1995 from the All-Russian Institute of the Scientific and Technical Information (VINITI), Russia. Since 1997, he is the head of the Natural Language and Text Processing Laboratory of the Computing Research Center, National Polytechnic Institute, Mexico City. He is an academician of Mexican Academy of Sciences since 2000 and National Researcher of Mexico (SNI) since 1998; author of about 300 publications on computational linguistics; see www.Gelbukh.com.*

Grigori Sidorov *was born in Moscow, Russia, in 1965. He obtained his Master degree in Structural and Applied Linguistics in 1988 from the Philological faculty of the "Lomonosov" Moscow State University, Russia, and his Ph.D. degree in Structural, Applied and Mathematical Linguistics in 1996 from the same faculty. Since 1998, he works for the Natural Language and Text Processing Laboratory of the Computing Research Center, National Polytechnic Institute, Mexico City. He is a National Researcher of Mexico (SNI) since 1999; author of about 100 publications on computational linguistics; see www.cic.ipn.mx/~sidorov.*

Adolfo Guzmán-Arenas *was born in Ixtaltepec, Oaxaca, México, in 1943. He obtained his Master degree in Electrical Engineering (Computer Science) in 1967 and his Ph.D. degree in Computer Science in 1968, both from MIT, USA, under supervision of Marvin L. Minsky. Since 1996, he works for the Computing Research Center, National Polytechnic Institute, Mexico, of which he was the Director in 1996–2002. He is laureate of various national an institutional awards, including the National Award in Sciences and Arts of Mexico, 1996. He is an ACM Fellow since 2002, academician of the Mexican Academy of Sciences, Mexican Academy of Engineering, and New-York Academy of Sciences; National Researcher of Mexico (SNI); author of numerous publications on computer science; see alum.mit.edu/www/aguzman.*