Exploring Large Document Collections with a Dynamic Hierarchy

Alexander Gelbukh and Grigori Sidorov

Center for Computing Research (CIC), National Polytechnic Institute (IPN),
Av. Juan Dios Batiz s/n, esq. Othon de Mendizabal, Zacatenco, CP 07738, Mexico D.F., MEXICO
*E-mail: gelbukh*gelbukh.com, sidorov*cic.ipn.mx*

Abstract. We present a method for dealing with the large document collections that takes into account the user needs in a better way. The user can choose the desired granularity of the hierarchy and then apply this hierarchy to document collection for classifying the documents. The granularity depends directly on the number of clusters. The clusters are presented to the user in two different ways: (1) as a representative document of the cluster; (2) as a set of keywords characterizing all documents of the given cluster.

Key Words: Document collection, dynamic hierarchy, document classification.

1. Introduction

Modern document collections contain huge amount of information. Usually it is impossible—and unnecessary—for the user to read at least titles of all documents in a large collection to get an idea of the kind of information it contains. There are two main ways to automate this process: hierarchical organization of the collection (as a tree of folders in the computer or the topic tree of Yahoo system) and search (as in Internet search engines such as Google).

However, both methods have their shortcomings. Hierarchical structures tend to be too rigid: the predefined organization of the data does not take into account the preferences of a specific user. This is improved in the search paradigm, see, for example, [2], [3], and [4]: it perfectly takes into account the interests of the specific user. However, it proves to be at the other extreme: it does not provide the user with any clue on what kind of information is there in the data collection or any guidance for the user not familiar with the specific thematic domain. In this paper we describe a system aimed at combining the advantages of both approaches: to give the user an insight into the data via a hierarchical structure while letting them adjust the way the data is partitioned in the hierarchy.

2. Suggested Method

We present the contents of the collection to the user as a set of clusters. The user can adjust various parameters of the clustering procedure, such as the desired number of clusters or the weights of individual terms used for clustering and reflecting the thematic interest profile of the specific user. Then the user identifies the clusters of his or her interest and "opens" them (using a metaphor of opening a subfolder in a folder tree). The contents of such subfolder are

Alexander Gelbukh, Grigori Sidorov. Exploring Large Document Collections with a Dynamic Hierarchy. International Conference INFORMATION-2004, Tokyo, Japan, November 2004

presented again as a set of clusters, which in their turn can be "opened." At each level, the parameters of clustering—most importantly, the number of clusters—can be adjusted individually; this constitutes the difference with a static hierarchy such as Yahoo topic tree.

At any stage the user can restrict the part of the collection under exploration or adjust the order in which the clusters are presented by formulating a query, like it is done in search engines. To restrict exploration to a part of the collection, Boolean queries are used; to order the clusters, vector space model is used [1]. In particular, the navigation method we suggest can be used to explore the search results returned by a search engine.

An interesting issue is the way a cluster can be presented to the user. If the user were presented with just a group of documents, this would not give any speedup in navigation as compared with just looking through all the documents in the collection. Accordingly, we present the clusters to the user in one of the following two ways: (1) as a representative document of the cluster; three variants of .selection of such a representative document are discussed in [5]; (2) as a set of keywords characterizing all documents of the given cluster.

The typical document, i.e., the most representative one, can be determined using some mathematical techniques, for example, we can choose a geometric center of the cluster, say, the document that has a least mean distance to all other cluster members.

There are different ways for choosing keywords. One of the ways is the direct choice of the most frequent words, obviously, not counting the stopwords. The other possibility is the thematic abstract, i.e., the main themes of the document are determined, see [6], and they are presented as a thematic abstract.

One of the modes of applying of the hierarchical dictionary is generalization in document collections: for example, if a document mentions *crocodiles*, *cows* and *dogs*, while another one contains words like *carburetors*, *wheels* and *engines*, the theme of the first document will be *animals* and of the second one – *cars*. Nevertheless, the possibility of such generation depends on the structure of the collection itself.

3. Implementation and Experimental Results

We implemented the above ideas in a prototype system. For better clustering, we used an ontology to determine the similarity between the words of different documents: two documents were considered similar if they have many words in common or have many synonymous words or direct or indirect co-hyponyms [6]. We also applied various linguistic techniques such as word sense disambiguation [7].

In the experiments with the resulting system we compared the time necessary for a user to familiarize himself or herself with a previously unseen document collection using our system, unstructured list of the documents, a search engine, or a static hierarchy (with a fixed small number of clusters). The users achieved better results with our system.

Scenario of the usage of the system is the following:

The user starts his initial search, for example: “computer”.

The system presents the feedback: “About 10 thousand documents were found”, in the same way that the modern search engines do but also adds the information like the following:

“These are divided in 5 groups:

1 group: keywords: nets, protocolos, HTML,

2 group: keywords: genetic algorithms, logic inference, Lisp,

3 group: keywords: natural language, text, syntactic ambiguity,

etc.”

The user chooses one group (for example, group 3).

All the process is repeated: The system responds: “The group 3 consists of 2 thousand documents, that are divided in 4 groups...”; the user chooses one smaller group, etc.

Note that the system works intelligently because it constructs the groups automatically on the basis of the linguistic analysis of the documents’ contents.

Also note that this division is fixed, i.e., once the hierarchy is constructed, it is presented to the user in the same way as the hierarchical structures of Yahoo. The approach that constructs the hierarchical structures beforehand is easier from the technical point of view but it may be inadequate for a user.

The clustering algorithms used by the program have different parameters that should be determined by the user (say, the desired number of clusters or the lexical-thematic aspect of comparison, etc.). Though there is a possibility to assign default values, they can suit very poorly to a specific user.

So, the user has an opportunity to experiment with various parameters interactively, i.e., the user can try different values of the parameters and verify their influence to the clustering results. Different possible modes of interaction with the user and the modes of presentation of parameters are themes of future investigations.

Applying this technique, the user can choose the specific division of the current subset of documents. Note that in this case the hierarchies generated for the same set of documents are not fixed, instead they are generated each time in the different ways. This is the reason why we call them dynamic hierarchies.

4. Conclusion

We have suggested a method of exploring large document collections, which combines the flexibility of search engines with the guidance provided by hierarchical ordering of the documents. The method has been implemented in a prototype system. The users achieve better results in exploration of previously unseen collections using our system than using baseline methods.

Acknowledgements

The work was done under partial support of Mexican Government (CONACyT, SNI) and National Polytechnic Institute, Mexico (CGPI, COFAA).

References

[1] Baeza-Yates, R., B. Ribero-Neto. *Modern Information Retrieval*. Addison Wesley, 1999.

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

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

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

[5] Gelbukh, A., M. Alexandrov, A. Bourek, P. Makagonov. "*Selection of Representative Documents for Clusters in a Document Collection*."| NLDB-2003, Lecture Notes in Informatics, Bonner Köllen Verlag, 2003, pp. 120–126.

[6] Gelbukh A., G. Sidorov, A. Guzman-Arenas. "*Use of a weighted topic hierarchy for document classification*." In: V. Matoušek et al. (Eds.). Text, Speech and Dialogue (TSD-99), Lecture Notes in Artificial Intelligence, N 1692, Springer-Verlag, pp. 130–135.

[7] Ledo Mezquita, Y., G. Sidorov, A. Gelbukh. "*Tool for Computer-Aided Spanish Word Sense Disambiguation*." In: Computational Linguistics and Intelligent Text Processing (CICLing-2003), Lecture Notes in Computer Science, N 2588, Springer-Verlag, pp. 277–280.