Text Categorization Using a Hierarchical Topic Dictionary

Alexander Gelbukh
Grigori Sidorov
Adolfo Guzmán-Arenas

{gelbukh,sidorov,aguzman}@pollux.cic.ipn.mx
Natural Language Laboratory, Center for Computing Research (CIC),
National Polytechnic Institute (IPN). Av. Juan Días Bátiz, Zacatenco, 07738 DF.
Mexico

Abstract

A statistical method of text categorization driven by a hierarchical topic dictionary is proposed. The method uses a dictionary with a simple structure and is insensible to inaccuracies in the dictionary; the dictionary is easily trainable on a manually classified document collection and even automatically translatable into different languages. A common sense-complaint way of assignment of the weights to the topics is discussed. The discussion is based on the experience with the system CLASSIFIER developed on the base of these methods.

1 Introduction[*]

We consider the task of text categorization by the topic of the document: for example, some documents are about *animals*, and some about *industry*. In this paper we consider the list of topics to be large but fixed. Our algorithm does not obtain the topics 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.

A problem arises of the optimal, or reasonable, degree of detail for such classification. For example, when classifying the Internet news for an "average" reader, the categories like *animals* or *industry* are quite appropriate, while for classification 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 classify the documents by the topics such as *mammals*, *herptiles*, *crustaceans*, etc.

[*] The work done under partial support of DEPI-IPN, CONACyT (26424-A), REDII-CONACyT, and SNI, Mexico.

In this paper, we will discuss the structure of the topic dictionary, the choice and use of the weights of individual nodes in the hierarchy, and some practical aspects of compilation of the topic dictionary.

2 Topic hierarchy and classification algorithm

In [Guzmán-Arenas, 1997; 1998] it was proposed to use a hierarchical dictionary for determining the main themes of a document. Technically, the dictionary consists of two parts: *keyword groups* representing individual topics, and a *hierarchy* of such topics.

A keyword group is a list of words or expressions related to the situation referred to by the name of the topic. For example, the topic *religion* lists 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.

The topic tree organizes the topics, as integral units, into a hierarchy or, more generally, a lattice (since some topics can belong to several nodes of the hierarchy).

The algorithm of application of the dictionary to the task of topic detection also consists of two parts: individual (leaf) topic detection and propagation of the topic weights up the tree.

The first part of the algorithm is responsible for detection individual (leaf) topics. Effectively, it answers, topic by topic, the following question: To what degree this document corresponds to the given topic? Such a question is answered for each topic individually. For more information on how the topic weights are determined (in a slightly different situation), see [Alexandrov and Gelbukh, 1999]. In the simplest case, the weight of a topic is the number (frequency) of words from the corresponding word list, found in the document.

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

The question discussed in the paper is how far are to be propagated the weights up to the tree, for the determined main topic of the document not to be trivially general, like *objects*.

3 Relevance and discrimination weights

Instead of simple word lists, some numeric weights can be used by the algorithm to define (1) the quantitative measures of relevance of the words for topics and (2) the measure of importance of the nodes of the hierarchy.

The first type of weights, which we call *relevance weights*, is associated with the links between words and topics and the links between the nodes in the tree. For example, if the document mentions the word *carburetor*, is it about *cars*? And the word *wheel*? I.e., how relevant is the word *carburetor* or *wheel* for the topic *cars*, how strong is their relationship? Intuitively, the contribution of the word *carburetor* into the topic *cars* is greater than that of the word *wheel*; thus, the link between *wheel* and *cars* is assigned a less weight.

It can be shown that the weight w_k^j of such a link (between a word *k* and a topic *j*, or between a topic *k* and its parent topic *j* in the tree) can defined as the mean relevance for the given topic of the documents containing this word: $w_k^j = \sum_{i \in D} r_i^j n_i^k \Big/ \sum_{i \in D} n_i^k$. Here the summation is done by all the available documents *D*, 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*.

Unfortunately, we are not aware of any reliable algorithm to automatically detect the measure of relevance r_i^j of the documents for the domains in an independent way. For the moment, such a measure is estimated manually by the expert, and then the system is trained on the set of documents. Alternatively, the expert can usually intuitively assign the relevance weights to the documents.

Both these approaches require manual work. To avoid it, as a practical approximation, for narrow enough themes the hypothesis can be assumed that the texts on this topic almost never occur in general texts (newspaper mixture). Then the expression for the weights can be simplified: $w_k^j = 1 \Big/ \sum_{i \in D} n_i^k$.

The main requirement for the second type of weights – the *discrimanation weights* – is their *discrimination power*: a topic should correspond to a (considerable) *subset* of documents. On the other hand, the topics that correspond to nearly all the documents in the data base are useless because they do not permit to make any relevan conclusions about the corresponding documents.

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 measure it as the dispersion: $w^j = \sum_{i \in D} \left(r_i^j - M\right)^2$, where $M = \sum_{i \in D} r_i^j \Big/ |D|$ is the average value of r_i^j over the current database *D*, and r_i^j is determined by the former algorithm, i.e., without taking into account the value of w^j . In a more precise manner, the information theory can be applied to the calculation of the weights; we will not discuss here this idea.

With this approach, for, say, a biological database, the weight of the topics like *animals, living things, nature* is low because all the documents equally mention these topics. On the other hand, for the newspaper mixture their weight is high, since many documents in it do not correspond to these topics, but still some considerable part do.

References

[Alexandrov and Gelbukh, 1999] M. Alexandrov and A. Gelbukh. Measures for Determining Thematic Structure of Documents with Domain Dictionaries. In *Proc. Workshop on Text Mining, International Joint Conference on Artificial Intelligence IJCAI-99,* to appear, 1999.

[Guzmán-Arenas, 1997] Adolfo Guzmán-Arenas. Hallando los temas principales en un artículo en español (in Spanish). *Soluciones Avanzadas,* 5(45):58, 5(49):66, 1997.

[Guzmán-Arenas, 1998] Adolfo Guzmán-Arenas. Finding the main themes in a Spanish document. Journal Expert Systems with Applications, 4(1/2):139-148,. January-February 1998.

[Gelbukh et al., 1997] Alexander Gelbukh, Igor Bolshakov, Sofía Galicia Haro. Automatic Learning of a Syntactical Government Patterns Dictionary from Web-Retrieved Texts. In *Proceedings of the International Conference on Automatic Learning and Discovery*, Pittsburgh, PA, USA, June 1998. Carnegie Mellon University.