Multiple label text categorization on a hierarchical thesaurus[⋆]

Francisco J. Ribadas, Erica Lloves, and Victor M. Darriba

Dept. de Informática, Univ. of Vigo
Campus de As Lagoas, s/n, 32004
Ourense, Spain
`ribadas@uvigo.es, darriba@uvigo.es`

Telémaco, I.D.S., S.L.
Parque Tecnológico de Galicia
Ourense, Spain
`ericalloves@gmail.com`

In many document processing tasks a correct identification of relevant topics offers a helpful starting point to develop advanced applications for browsing and searching in large collections of documents. In this context, one of the most valuable tools are specialized thesauri. These kinds of structure organize a set of concepts relevant to a given domain in a hierarchical structure, making it possible to employ a sort of controlled vocabulary to simplify document processing.

In this paper we describe our work on the automatic association of relevant topics, taken from a structured thesaurus, to documents written in natural languages. In our case we are interested in the domain of legislative texts in Spanish. We have an available thesaurus, manually built, with more than 160 concepts, arranged in a tree structure. We also have a collection of about 10,000 legislative documents, whose main topics have been identified by humans according to the entries in that thesaurus. The approach we have followed models thesaurus topic assignment as a multiple label classification problem, where the whole set of possible classes is hierarchically organized. Many previous proposals [2] have dealt with text categorization, but the case of hierarchical classes is usually omited or the generalization to multiple label classification is not directly supported.

The strategy we have chosen is inspired by Koller and Sahami's work [1] and takes advantage of the class hierarchy to simplify the classification task in two aspects. Firstly, the global classification problem is reduced to a sequence of partial classifications, guided by the structure of our topic tree. Secondly, the computational cost for each classification step is reduced and the resulting quality is improved by means of the use of a specific set of features, exclusive to each node in the hierarchy. The main difference with Koller and Sahami's proposal is the final output of our classifier. Our aim is to get a set of relevant topics taken from the thesaurus, ordered according to their relevance, instead of a single class. We replace the greedy approach used in the original work, where only the best successor for each node was considered at each step. In our proposal, we proceed level by level, and all the paths starting at successors with high evaluation values are taken into account, and they are followed until no further expansion is admissible. The final set of topics is composed of those class

[⋆] Research partially supported by Telemaco, Información, Documentación y Sistemas, SL, Xunta de Galicia(PGIDIT05SIN044E, PGIDIT05PXIC30501PN, *Galician Network for Language Processing and Information Retrieval*), Ministry of Education and Science (TIN2004-07246-C03-01) and University of Vigo

nodes, ranking them according to the strength of the classification steps that lead to them. Our proposal can be outlined in the following main points.

Document preprocessing. Since this is a first approach to this problem, we have tried to avoid using complex linguistic resources, like taggers or lemmatizers. The first step was to extract plain text from original HTML and PDF documents. From those text files we select their content words. To omit non-relevant words, we use a generic stop-word list for Spanish and a set of specific stop-words for legislative domains. Remaining words are normalized using stemming rules to overcome lexical variation problems. All of the resulting stems will be taken into account as potential features, suitable for describing the document.

Training phase. In this phase we take the whole training collection with the set of topics associated to each document and we traverse the topic tree, performing two tasks at every thesaurus entry. For each node, we select the set of documents with at least one associated topic that is a descendant of the current concept. With these documents we apply simple feature selection techniques to find a set of features with the highest discrimination power among the different branches starting at this node. Then, a specialized classifier is trained to select the most promising branches at current level. For each document a feature vector is built. Only the relevant stems selected for the current concept are employed, using their *tf-idf* as feature values. The class for this training vector will be the current topic, if it is actually associated with the document, or the label of one of its sons, if some topic associated to the document falls into that branch. We have employed WEKA Machine Learning Engine [3] to train the specialized classifiers for each topic in our hierarchical thesaurus. We have tested different classification algorithms to be used inside this hierarchical classification scheme, obtaining the most promising results with k Nearest Neighbors (k-NN) learning approaches.

Classification phase. Once all of the partial classifiers have been trained, the assignment of topics to new documents means traversing the thesaurus tree. Starting at its root, the feature vector for the document is built using the selected features for each node, and the most promising branches according to the classificator results are followed. The routing decisions taken at each node are controlled by simple thresholds that take into account both the number of alternatives at each node and the strength of the potential classes. If the class for the current topic is higher than this threshold it is considered to be suitable as a topic for this document. If no successor classes have sufficient strength, the deeping is stopped. The final list of potential topics is ranked acording to the set of values obtained in the sequence of partial classifications that lead to them.

References

1. D. Koller and M. Sahami. Hierarchically classifying documents using very few words. In *Proc. of 14th Int. Conf. on Machine Learning*, pp. 170–178, Nashville, US, 1997
2. F. Sebastiani. Machine learning in automated text categorization. In *ACM Computing Surveys, 24-1*, pp. 1–47, 2002
3. I. Witten and E. Frank Data Mining: Practical machine learning tools and techniques, 2nd Ed. *Morgan Kaufmann*, San Francisco, 2005.