International Journal of Approximate Reasoning 50 (2009) 932–944

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

journal homepage: www.elsevier.com/locate/ijar

Bayesian network models for hierarchical text classification from a thesaurus

Luis M. de Campos, Alfonso E. Romero *

Departamento de Ciencias de la Computación e Inteligencia Artificial, E.T.S.I. Informática y de Telecomunicación, Universidad de Granada, Daniel Saucedo Aranda, s/n, 18071 Granada, Spain

ARTICLE INFO

Article history:
Received 18 March 2008
Received in revised form 12 September 2008
Accepted 28 October 2008
Available online 27 November 2008

Keywords:
Bayesian networks
Document categorization
Hierarchical classification
Thesauri

ABSTRACT

We propose a method which, given a document to be classified, automatically generates an ordered set of appropriate descriptors extracted from a thesaurus. The method creates a Bayesian network to model the thesaurus and uses probabilistic inference to select the set of descriptors having high posterior probability of being relevant given the available evidence (the document to be classified). Our model can be used without having preclassified training documents, although it improves its performance as long as more training data become available. We have tested the classification model using a document dataset containing parliamentary resolutions from the regional Parliament of Andalucía at Spain, which were manually indexed from the Eurovoc thesaurus, also carrying out an experimental comparison with other standard text classifiers.

1. Introduction

To improve organizational aspects and facilitate fast access to relevant information relative to a particular subject, document collections from many organizations are classified according to their content using a set of descriptors extracted from some kind of controlled vocabulary or thesaurus. For example, most of the parliaments in Europe use a thesaurus called Eurovoc to classify parliamentary resolutions, the Food and Agricultural Organization employs Agrovoc to categorize its documents, several organizations use the National Agriculture Library Thesaurus (NALT), and the National Library of Medicine uses MeSH to index articles from biomedical journals. The process of assigning descriptors in the thesaurus to the documents is almost always carried out manually by a team of expert documentalists. The objective of this work is the development of a computerized tool to assist the human experts in this process. We believe that it is not realistic to try to design a completely automatic classification process, given the critical nature of the classification task in many contexts, and final human supervision will always be required in real environments.

The scope of our research is therefore automatic subject indexing from a controlled vocabulary [8,17] and hierarchical text classification [18,21]. There are several characteristics in this problem which make it difficult: (1) as each descriptor in the thesaurus represents a different class/category, it is a problem of high dimensionality (we are managing several thousand descriptors); (2) it is also a multi-label problem, because a document may be associated with several classes, exhibiting also a high variability in the number of descriptors being assigned to each document[1]; (3) there are explicit (hierarchical) relationships between the class labels, so that they are not independent among each other; (4) the training data can be quite unbalanced, having a very different number of documents associated to each class.

* Corresponding author.
E-mail addresses: lci@decsai.ugr.es (L.M. de Campos), aeromero@decsai.ugr.es (A.E. Romero).

[1] Between 1 and 14 in the document collection used in the experiments.

0888-613X/$ - see front matter Crown Copyright © 2008 Published by Elsevier Inc. All rights reserved.
doi:10.1016/j.ijar.2008.10.006

An important characteristic of the model that we are going to propose is that no training is required to start using the system. Initially we shall exploit only the hierarchical and lexical information from the thesaurus to build the classifier. This is an advantage because the model may be used with almost any thesaurus and without having preclassified documents (in a large hierarchy, the amount of preclassified documents necessary for training may be huge). On the other hand, this is also a weakness because any kind of information not considered in the thesaurus (e.g. other relations, specific information handled by documentalists,...) will not be taken into account and, therefore, we should not expect very high success rates in comparison with classifiers that are built using training data [4,7,14,20]. In this sense our initial proposal is more similar to the work in [1,2], where a method to populate an initially empty taxonomy is proposed. The working hypothesis is that a documentalist would prefer to confirm or discard a given classification hypothesis proposed by the system rather than examining all the possible alternatives.

Nevertheless, the proposed model can also naturally incorporate training data in order to improve its performance: The information provided by preclassified documents can be appropriately merged with the hierarchical and equivalence relationships among the descriptors in the thesaurus, in order to obtain a classifier better than the one we would obtain by using only the training documents.

Another important characteristic of our model is that is based on Bayesian networks. To the best of our knowledge, no Bayesian network-based models other than naive Bayes have been proposed to deal with this kind of problems [12]. We create a Bayesian network to model the hierarchical and equivalence relationships in the thesaurus, and next we extend it to also use training data. Then, given a document to be classified, its terms are instantiated in the network and a probabilistic inference algorithm, especifically designed and particularly efficient, computes the posterior probabilities of the descriptors in the thesaurus.

The paper is organized as follows: In Section 2 we describe the proposed Bayesian network[2] model of a thesaurus, whereas the extension of the model to cope with training data is described in Section 3. The experimental evaluation is explained in Section 4. Finally, Section 5 contains the final remarks and some proposals for future work.

2. The Bayesian network representing a thesaurus

In this section we shall first introduce basic notions relative to the composition and structure of a thesaurus; next, we describe the Bayesian network model proposed to represent it, including the graphical structure, the conditional probabilities and the inference mechanism.

2.1. Thesauri

Broadly speaking, a thesaurus consists of a set of terms, which are relevant to a certain domain of knowledge, and a set of semantic relationships between them. The basic units of a thesaurus are *descriptors* or *indexing terms*, which are words or expressions which denote in unambiguous fashion the constituent concepts of the field covered by the thesaurus. A thesaurus also comprises *non-descriptors* or *entry terms*, which are words or expressions that denote the same or a more or less equivalent concept as a descriptor in the language of the thesaurus. The three most common types of semantic relationships are equivalence, hierarchical and associative relationships.

The *equivalence relationship* between descriptors and non-descriptors may cover relationships of several types: genuine synonymy, near-synonymy, antonymy and inclusion, when a descriptor embraces one or more specific concepts which are given the status of non-descriptors because they are not often used. It is usually represented by the abbreviations "UF" (Used For), between the descriptor and the non-descriptor(s) it represents, and "USE" between a non-descriptor and the descriptor which takes its place. The *hierarchical relationship* between descriptors is shown by the abbreviations: "BT" (Broader Term) between a specific descriptor and a more generic descriptor, and its dual "NT" (Narrower Term) between a generic descriptor and a more specific descriptor. Descriptors which do not contain other more specific descriptors are called *basic descriptors*; otherwise they are called *complex descriptors*. Descriptors which are not contained in any other broader descriptors are *top descriptors*. Sometimes a few descriptors are polyhierarchical (they have more than one broader descriptor), which means that the hierarchical relationships may form a graph instead of a tree. The *associative relationship*, shown by the abbreviation "RT" (Related Term), relates two descriptors that do not meet the criteria for an equivalence nor a hierarchical relationship. It is used to suggest another descriptor that would be helpful for the thesaurus user to search by. In this work we shall not consider associative relationships.

2.1.1. Example

Eurovoc is a multilingual thesaurus that provides a means of indexing the documents in the documentation systems of the European institutions and of their users. Fig. 1 displays the BT relationships between some descriptors of Eurovoc and the USE relationships between the non-descriptors and these descriptors.[3] There are two complex descriptors, *health service* and *health policy*, and three basic descriptors, *medical centre*, *medical institution* and *psychiatric institution*. *Health service* is the broad-

[2] We assume that the reader has at least a basic background on Bayesian networks.

[3] The English version of Eurovoc comprises 6645 descriptors and 6769 non-descriptors, together with 6669 BT/NT relationships.