Sense Cluster based Categorization and Clustering of Abstracts

Davide Buscaldi[1], Paolo Rosso[1], Mikhail Alexandrov[2], and Alfons Juan Ciscar[1]

[1] Dpto. Sistemas Informáticos y Computación (DSIC),
Universidad Politécnica de Valencia, Spain
{dbuscaldi, prosso, ajuan}@dsic.upv.es
[2] Center for Computing Research,
National Polytechnic Institute, Mexico
dyner1950@mail.ru

Abstract. This paper focuses on the use of sense clusters for classification and clustering of very short texts such as conference abstracts. Common keyword-based techniques are effective for very short documents only when the data pertain to different domains. In the case of conference abstracts, all the documents are from a narrow domain (i.e., share a similar terminology), that increases the difficulty of the task. Sense clusters are extracted from abstracts, exploiting the WordNet relationships existing between words in the same text. Experiments were carried out both for the categorization task, using Bernoulli mixtures for binary data, and the clustering task, by means of Stein's MajorClust method.

1 Introduction

Typical approaches to document clustering and categorization in a given domain are to transform the textual documents into vector form, by using a list of index keywords. This kind of approaches has also been used for clustering etherogeneous short documents (e.g. documents containing 50-100 words) with good results. However, term-based approaches usually give unstable or imprecise results when applied to documents from one narrow domain.

Previous works on narrow-domain short document classification obtained good results by using supervised methods and set of keywords (*itemsets*) as index terms [3].

In this work, we exploited the linguistic information extracted from WordNet in order to extract key *concept clusters* from the documents, using the method proposed by Bo-Yeong Kang *et al.* [5], which is based on semantic relationships between the terms in the document. Concept clusters are used as index words.

Various methods have been tested for the categorization and clustering task, including Bernoulli mixture models, which have been investigated for text categorization in [4]. Text categorization procedures are based on either binary or integer-valued features. In our case, due to the low absolute frequency observable in short documents, we used only the information if an index term was or not in the abstract, thus obtaining a binary representation of each document.