A Technique for Automatic Generation of Rule-based Text Classifiers exploiting Negative Information

Chiara Cumbo

Dipartimento di Matematica,
Università della Calabria
87036 *Rende, Italy*
email : cumbo@mat.unical.it

Abstract

Information Retrieval is concerned with locating information that will satisfy a user's information need. Traditionally, the emphasis has been on text retrieval: providing access to natural language texts where the set of documents to be searched is large and topically diverse. In a text categorization task, the system is responsible for assigning a document to one or more categories from among a given set of categories.

In this way, users are allowed to browse more easily the set of texts of their own interests, by navigating in category hierarchies. This paradigm is very effective for retrieval and for filtering of information but also in the development of user-driven on-line services. Given the large amounts of documents involved in the above applications, automated approaches to categorize data efficiently are needed.

The automated categorization (or classification) of texts into prespecified categories, although dating back to the early '60s, has witnessed a booming interest in the last ten years, due to the increased availability of documents in digital form and the ensuing need to organize them. In the research community the dominant approach to this problem is based on the application of supervised machine learning techniques: a general inductive process automatically builds a classifier by learning, from a set of previously classified documents, the characteristics of one or more categories. The advantages of this approach over the knowledge engineering approach (consisting in the manual definition of a classifier by domain experts) are a very good effectiveness, considerable savings in terms of expert manpower, and straightforward portability to different domains.

A good text classifier is a classifier that efficiently categorizes large sets of text documents in a reasonable time frame and with an acceptable accuracy, and that provides classification rules that are human readable for possible fine-tuning. If the training of the classifier is also quick, this could become in some application domains a good asset for the classifier. Many techniques and algorithms for automatic text categorization have been devised. According to published literature, some are more accurate than others, and some provide more interpretable classification models than others. However, none can combine all the beneficial properties enumerated above.

In this dissertation, we have defined and implemented a novel approach OLEX,

for automatic text categorization. OLEX relies on an optimization algorithm whereby a set of both positive and negative information are generated from a set of training documents in order to learn profiles of predefined categories with respect to which we wish to construct text classifiers. The proposed method is simple and elegant. Despite this our text categorization method proves to be efficient and effective, and experiments on well-known collections show that the classifier performs well. In addition, training as well as classification are both fast and the generated rules are human readable.

OLEX has been fully integrated into an industrial text classification system developed at Exeura s.r.l.

Briefly, the main contributions of the thesis are the following:

1. We study methods and systems for automatic text classification, analyze their complexity and their exploitation for a critical comparison.
2. We design a new machine learning method for generating logic rules for text categorization.
3. We implement our approach in a prototype, the OLEX system.
4. We perform a systematic experimentation and report experimental results on a number of well-known benchmark text collections to assess the impact of our approach and to compare it with respect to other systems.
5. We integrate the support for learning process in OLEX Content Management Suite, within project PIA-Exeura-03-06.