Supervised Term Weighting for Automated Text Categorization

by Franca Debole and Fabrizio Sebastiani

Researchers from ISTI-CNR, Pisa, aim at producing better text classification methods through the use of supervised learning techniques in the generation of the internal representations of the texts.

Text classification (TC) is the activity of automatically building, by means of machine learning (ML) techniques, programs ('text classifiers') capable of 'pigeonholing' natural language texts, ie placing them in categories from a predefined set, according to an analysis of their content. Instances of text classification are topic spotting, spam filtering, genre classification, or automated authorship attribution, depending on the nature and meaning of the categories being considered.

The construction of an automatic text classifier is usually articulated in three phases. The first phase is term selection, in which the most relevant terms for the classification task are identified. The second phase is term weighting, in which document-dependent weights for the selected terms are computed so as to generate a vectorial representation for each document, in which each term is weighted by its contribution to the extensional semantics of the document. The third phase is classifier learning, in which a learning device generates a classifier from the vectorial representations of the training documents.

This entire process involves an activity of 'supervised learning', ie one in which information on the membership of training documents in specific categories is used. Traditionally, supervised learning only enters into phases 1 and 3; phase 2, instead, usually relies on techniques borrowed from text search such as tf-idf ('text frequency * inverse document frequency'), a weighting function based on the distribution of the terms within the document and within the collection, where a high value indicates that the word occurs often in the document and does not occur in many other documents. As a consequence, these techniques do not exploit the information provided by training documents since text search does not involve any training documents.

In our current work we propose that learning from training documents should also affect the term weighting phase, ie that information on the membership of training documents in specific categories be used to determine term weights. We call this idea supervised term weighting (STW). As an example of STW we propose a number of 'supervised variants' of tf*idf weighting, obtained by replacing the idf part of tf*idf with the same function that has previously been used in the term selection phase. The rationale of replacing idf lies in the fact that idf represents a measure of the document-independent value of a term, but as such it is suboptimal in a text classification context, in the sense that it relies on an intuition ("the document-independent value of a term is inversely proportional to the number of documents in which it occurs") that is valid also in information retrieval tasks in which no training data are available. Feature selection functions rely instead on an intuition ('the document-independent value of a term is directly proportional to how differently the term is distributed in the positive and negative examples of the categories of interest') that refers to the presence of categories, and that is thus specific to tasks in which category data is available.

We have run STW experiments on Reuters-21578, the standard benchmark of text classification research, with three classifier learning methods (Rocchio, kNN, support vector machines), three term selection functions (information gain, chi-square, and gain ratio), and two policies for addressing term selection and weighting ("local" and "global"). Results show that STW is a powerful notion since, when instantiated with the 'gain ratio' feature selection function, it can bring about improvements as high as 11% in accuracy over a standard tf*idf representation with a support

vector machine learner.

Links:
http://faure.iei.pi.cnr.it/~fabrizio/Publications/SAC03b.pdf

Please contact:
Fabrizio Sebastiani, ISTI-CNR
Tel: +39 050 3152892
E-mail: fabrizio.sebastiani@isti.cnr.it