With or without context

Automatic text categorization using semantic kernels

Johan Eklund

Academic dissertation for the Degree of Doctor of Philosophy in Library and Information Science at the University of Borås to be publicly defended on Friday 15 April at 13.00 in lecture room C203, the University of Borås, Allégatan 1, Borås

The Swedish School of Library and Information Science
The University of Borås

Title: With or without context: Automatic text categorization using semantic kernels

Language: English, with a summary in Swedish

Available: http://urn.kb.se/resolve?urn=urn:nbn:se:hb:diva-8949
ISBN (printed version) 978-91-981654-8-7
ISBN (digital version) 978-91-981654-9-4
ISSN 1103-6990

In this thesis text categorization is investigated in four dimensions of analysis: theoretically as well as empirically, and as a manual as well as a machine-based process. In the first four chapters we look at the theoretical foundation of subject classification of text documents, with a certain focus on classification as as a procedure for organizing documents in libraries. A working hypothesis used in the theoretical analysis is that classification of documents is a process that involves translations between statements in different languages, both natural and artificial. We further investigate the close relationships between structures in classification languages and the order relations and topological structures that arise from classification.

A classification algorithm that gets a special focus in the subsequent chapters is the *support vector machine* (SVM), which in its original formulation is a binary classifier in linear vector spaces, but has been extended to handle classification problems for which the categories are not linearly separable. To this end the algorithm utilizes a category of functions called *kernels*, which induce feature spaces by means of high-dimensional and often non-linear maps. For the empirical part of this study we investigate the classification performance of semantic kernels generated by different measures of semantic similarity. One category of such measures is based on the *latent semantic analysis* and the *random indexing* methods, which generates term vectors by using co-occurrence data from text collections. Another semantic measure used in this study is *pointwise mutual information*. In addition to the empirical study of semantic kernels we also investigate the performance of a term weighting scheme called *divergence from randomness*, that has hitherto received little attention within the area of automatic text categorization.

The result of the empirical part of this study shows that the semantic kernels generally outperform the “standard” (non-semantic) linear kernel, especially for small training sets. A conclusion that can be drawn with respect to the investigated datasets is therefore that semantic information in the kernel in general improves its classification performance, and that the difference between the standard kernel and the semantic kernels is particularly large for small training sets. Another clear trend in the result is that the divergence from randomness weighting scheme yields a classification performance surpassing that of the common *tf-idf* weighting scheme.

Keywords: automatic text categorization, subject classification, machine learning, computational linguistics, support vector machines, semantic kernels, term weighting, divergence from randomness