A simple base–line text-categorizer for evaluating the effect of feature extraction in text mining applications

Günter Neumann and Michael Kappes

German Research Center for Artificial Intelligence (DFKI)
LT–Lab, DFKI Saarbrücken, D-66123 Saarbrücken, Germany

Abstract. Text Mining (TM) is concerned with the task of extracting relevant information from natural language (NL) text documents and to search for interesting relationships between the extracted features. A challenging feature of TM systems is that the information is only implicitly encoded in an unstructured way in NL texts. Thus, a major first step in every TM system is to map the NL text to a structured internal representation, which is then processed usually by known data mining algorithms. Basically, NL–feature recognition methods might range from simple word stemming algorithm to complex syntactic or semantically based ones. Our major scientific question is: what will be the impact of the different sort of extracted features on the data mining algorithm? We have chosen Text Categorization (TC) as our initial TM application. Since the standard TC algorithms are not very transparent wrt. the parametrization and adaptation to different feature extraction methods (see also Neumann&Schmeier 2002), we decided to develop our own very simple base–line categorizer called SimpleCat (SC).

Documents and categories are considered as consisting of a set of independent words. For each category, SC creates two data structures (category profiles): surface-focused index words with a high discrimination value and semantics–oriented topic words. Having two different word lists will allow as to apply different feature extraction methods and to evaluate their effect on the mining problem. Document-pivoted classification is done by first creating two category rankings (each for index and topic words) which are combined to one ranking (m-ary classifier). Category-pivoted classifying means to extract the documents of a particular category.

We executed tests on three corpora in order to evaluate the performance of the base–line algorithm. So far, no preprocessing other than recognition of simple word borders (e.g., space) are used. Optimization criterion is the micro-averaged F1-measure (miF1). We used Reuters-21578 corpus to compare SC with other methods. SC's miF1 is 78.84% (recall=71.18%, precision=88.34%). For comparison (Yang&Liu 1999): NaiveBayes miF1=79.56%, Support-Vector-Machines miF1=85.99%. SC's data pool simply consisted of the index lists and 125 topic words per category. Tests on Bundespresseamt-corpus (8817 German news wire stories, 16 categories) show that SC is able to support document distribution systems. A corpus consisting of approx. 8000 medical abstracts (39 categories) was used to compare classification results between German and English texts. Although SC uses very simple methods (basically maximum likelihood) it performed well enough as a base–line. Next, we hope to improve the performance by using the different linguistic methods developed at our Lab for feature extraction.

References

NEUMANN, G. and S. SCHMEIER (2002): Shallow Natural Language Technology and Text Mining. In: T. Joachims, E. Leopold (Eds.): KI, German Journal on Artificial Intelligence, *Special Issue on Text Mining*, T. Joachims, E. Leopold (Eds.), to appear.

YANG, Y. and X. Liu (1999): *A re–examination of text categorization methods.* In: Proc. of SIGIR–99, Berkely.

Keywords

Text mining, data structure analysis, text categorization