Handling Texts ? A Challenge for Data Mining

Katharina Morik

Technical University Dortmund
Dept. Computer Science VIII
44221 Dortmund (Germany)
katharina.morik@uni-dortmund.de

The amount of data in free form by far surpasses the structured records in databases in their number. However, standard learning algorithms require observations in the form of vectors given a fixed set of attributes. For texts, there is no such fixed set of attributes. The bag of words representation yields vectors with as many components as there are words in a language. Hence, the classification of documents represented as bag of word vectors demands efficient learning algorithms. The TCat model for the support vector machine (Joachims 2002) offers a sound performance estimation for text classification.

The huge mass of documents, in principle, offers answers to many questions and is one of the most important sources of knowledge. However, information retrieval and text classification deliver merely the document, in which the answer can be found by a human reader ? not the answer itself. Hence, information extraction has become an important topic: if we can extract information from text, we can apply standard machine learning to the extracted facts (Craven et al. 1998). First, information extraction has to recognize Named Entities (see, e.g., Roessler, Morik 2005). Second, relations between these become the nucleus of events. Extracting events from a complex web site with long documents allows to automatically discover regularities which are otherwise hidden in the mass of sentences (see, e.g., Jungermann, Morik 2008).

References

Craven, M., D. DiPasquo, D. Freitag, A. McCallum, T. Mitchell, K. Nigam, and S. Slattery (1998). Learning to extract knowledge from the world wide web. In *Proc. of the 1998 National Conference on Artificial Intelligence*.

Joachims, T. (2002). *Learning to Classify Text using Support Vector Machines*. Kluwer.

Jungermann, F. and K. Morik (2008). Enhanced services for targeted information retrieval by event extraction and data mining. In *Proc. of the 13th International Conference on Applications of Natural Language to Information Systems NLDB*.

Marc Roessler, K. M. (2005). Using unlabeled texts for named-entity recognition. In *Proc. of the ICML Workshop on Multiple View Learning*.