Feature Generation for Textual Information Retrieval Using World Knowledge

Evgeniy Gabrilovich
Department of Computer Science
Technion - Israel Institute of Technology
Technion City, Haifa 32000, Israel
gabr@cs.technion.ac.il
http://www.cs.technion.ac.il/~gabr/papers/phd-thesis.pdf

Imagine an automatic news filtering system that tracks company news. Given the news item "FDA approves ciprofloxacin for victims of anthrax inhalation", how can the system know that the drug mentioned is an antibiotic produced by Bayer? Or consider an information professional searching for data on RFID technology – how can a computer understand that the item "Wal-Mart supply chain goes real time" is relevant for the search? Algorithms we present can do just that.

When humans approach text processing tasks, such as text categorization, they interpret documents in the context of their background knowledge and experience. On the other hand, conventional information retrieval systems represent documents as bags of words, and are restricted to learning from individual word occurrences in the (necessarily limited) training set. We propose to enrich document representation through automatic use of vast repositories of human knowledge. To this end, we use Wikipedia and the Open Directory Project, the largest encyclopedia and Web directory, respectively. Wikipedia articles and ODP categories represent knowledge concepts. In the preprocessing phase, a feature generator analyzes the input documents and maps them onto relevant concepts. The latter give rise to a set of generated features that either augment or replace the standard bag of words. Feature generation is accomplished through contextual analysis of document text, thus implicitly performing word sense disambiguation. Coupled with the ability to generalize from words to concepts, this approach addresses the two main problems of natural language processing – synonymy and polysemy.

Categorizing documents with the aid of knowledge-based features leverages information that cannot be deduced from the training documents alone. Empirical results confirm that this knowledge-intensive representation brings text categorization to a qualitatively new level of performance across a diverse collection of datasets.

We also adapt our feature generation methodology for another task in natural language processing, namely, automatic assessment of semantic relatedness of words and texts. Previous state of the art results are based on Latent Semantic Analysis, which represents documents in the space of "latent concepts" computed using Singular Value Decomposition. We propose Explicit Semantic Analysis, which uses the feature generator methodology to represent the meaning of text fragments in a high-dimensional space of features based on natural concepts identified and described by humans. Computing semantic relatedness in this space yields substantial improvements, as judged by the very high correlation of computed scores with human judgments.