Text categorization using compression models

Eibe Frank, Chang Chui and Ian H. Witten

Department of Computer Science, University of Waikato, New Zealand
{eibe, ckc1, ihw}@cs.waikato.ac.nz

Text categorization is the assignment of natural language texts to predefined categories based on their content. The use of predefined categories implies a "supervised learning" approach to categorization, where already-classified articles—which effectively define the categories—are used as "training data" to build a model that can be used for classifying new articles that comprise the "test data." Typical approaches extract "features" from articles, and use the feature vectors as input to a machine learning scheme that learns how to classify articles. The features are generally words.

It has often been observed that compression seems to provide a very promising alternative approach to categorization. The overall compression of an article with respect to different models can be compared to see which one it fits most closely. Such a scheme has several potential advantages: it yields an overall judgement on the document as a whole, rather than discarding information by pre-selecting features; it avoids the messy and rather artificial problem of defining word boundaries; it deals uniformly with morphological variants of words; depending on the model (and its order), it can take account of phrasal effects that span word boundaries; it offers a uniform way of dealing with different types of documents—for example, arbitrary files in a computer system; it generally minimizes arbitrary decisions that inevitably need to be taken to render any learning scheme practical.

We have performed extensive experiments on the use of PPM compression models for categorization using the standard Reuters-21578 dataset. This has involved working out how to deal with the (normal) situation where a document may belong to several categories (not merely choosing the one that it fits best). We obtained encouraging results on two-category situations, and the results on the general problem seem reasonably impressive—in one case outstanding. PPM succeeds in categorizing the majority of documents correctly, and compares well with simple machine learning schemes. However, we find that it does not compete with the published state of the art in the use of machine learning for text categorization. PPM produces inferior results because it is insensitive to subtle differences between articles that belong to a category and those that do not. We do not believe our results are specific to PPM. If the occurrence of a single word determines whether an article belongs to a category or not (as it sometimes does), any compression scheme will likely fail to classify the article correctly. Machine learning schemes fare better because they automatically eliminate irrelevant features and concentrate on the most discriminating ones.

Frank, E., Chui, C., and Witten, I.H. (2000) "Text categorization using compression models." Technical Report 00/02, Department of Computer Science, University of Waikato.