A quantitative analysis of legal word use

Kees van Noortwijk
Centre for Computers and Law, Erasmus University Rotterdam
Email: vannoortwijk@frg.eur.nl

Abstract

An important task in legal knowledge management is the management of legal documents. Especially the retrieval of documents stored in electronic repositories (including commercial databanks) proves to be far more complicated than one would anticipate, at least when a certain level of accuracy is required.

This paper argues that legal document retrieval can be improved when more is known about the characteristics of the documents. Some of these characteristics, such as added 'metadata' describing the subject matter of documents, are already widely used by retrieval systems. One characteristic that is not commonly used, however, is the *word use* in these documents. Many people feel that the word use in legal documents seems to be different from that in other documents. If such a difference would indeed be present, that is if (certain types of) legal documents – statute law texts, case reports – would contain different words or would show different word use patterns than 'general' texts, such characteristics might be useful to improve retrieval systems.

A pilot study on this subject, performed for the Dutch language a decade ago[1], showed some interesting results. For instance, word frequency distributions proved to be quite different in legal texts than in 'general Dutch'. Now, two corpora containing thousands of legal documents (one containing statute law texts and one containing case reports) and a corpus containing general texts, all in British English, have been compiled. The corpora are of roughly equal sizes (around 16 million words each). The 'general' corpus consists of a random sample from the 'British National Corpus'[2], the two legal corpora consist of legal texts available on the internet. Cases for the case reports corpus have been selected in such a way that the percentage of Supreme Court, High Court, County Court etc. cases more or less corresponds to that in the ten year old Dutch case law corpus, which will facilitate inter-language comparison.

Using these corpora, research is now performed to map as precisely as possible the differences between word use in the respective language types. The results will be compared to those from the Dutch language study. Some first results from this are reported in this paper. A report containing the complete results will appear later this year.

It is expected that results from this research project will be of importance for the development of new, more 'intelligent' legal document retrieval systems in the near future. The final section of the paper contains examples of this.

---

[1] C. van Noortwijk, *Het woordgebruik meester* (Legal Word Use), with a summary in English, Lelystad: Vermande 1995, ISBN 90-5458-261-8

[2] The British National Corpus (BNC) is a 100 million word collection of samples of written and spoken language from a wide range of sources, designed to represent a wide cross-section of current British English. For this project, only the written sources were used. See http://www.natcorp.ox.ac.uk/.