Identifying Significant Topics of a Document

“Lead Inventor: Faye (Nina) P. Wacholder, PhD

Key Word Identification in Documents: Ambiguous: Statistical techniques for identifying keywords in a document traditionally rely on calculating the frequency of individual stems and/or words. However, stems are ambiguous ("trad" might refer to "trader" or "tradition"), as are the meanings of isolated words ("state" might mean a political entity or a mode of being). Thus, the resulting keyword lists often do not often accurately represent the aboutness of a document. Algorithms have since been developed to process proper nouns, as well as technical terms in scientific documents, but these algorithms are not domain-general. That is, they are not suited for identifying common noun phrases in an open-ended set of document types, particularly short articles.

Significant Topics Identified in Documents: The invention is a method for identifying significant topics of a document, comprised of the following steps: (1) extracting from a document a list of simplex noun phrases (NPs) and their corresponding heads, (2) clustering the simplex NPs by head, and (3) ranking the clustered simplex NPs by head in accordance with a significance measure. As simplex NPs may contain a determiner, an adjective, a noun, and no further speech elements such as a preposition or participial verb, the head, or noun, is inevitably the last word of the simplex NP in English. Simplex NPs, which are semantically and syntactically coherent, provide adequate content representation out of the context of the entire document, thus making the invention useful for various natural language processing applications.

Applications: • Automatic indexing of print or electronic texts • Summarization of document content • Keyword content with which to filter documents relevant for a specific query • Advanced information extraction where information about a specific topic from different parts of a document can be merged

Advantages: • Simplex noun phrases, in contrast with individual stems or words, are unambiguous • Simplex noun phrases are semantically and syntactically coherent • Domain-general method, which does not depend on domain-specific particulars of a document • Extensive applications in current information-rich, but knowledge-poor environment

Licensing Status: Available for Licensing and Sponsored Research Support”

Inventors

Judith Klavans