PHYS.ORG

Fast algorithm extracts, compares document meaning

September 25 2012

A computer program could compare two documents and work spot the differences in their meaning using a fast semantic algorithm developed by information scientists in Poland.

Writing in the *International Journal of Intelligent Information and Database Systems*, Andrzej Sieminski of the Technical University of Wroclaw, explains that extracting meaning and calculating the level of semantic similarity between two pieces of texts is a very difficult task, without human intervention. There have been various methods proposed by computer scientists for addressing this problem, but they all suffer from computational complexity, he says.

Sieminski has now attempted to reduce this complexity by merging a computationally efficient statistical approach to text analysis with a semantic component. Tests of the algorithm on English and Polish tests work well. The test set consisted of 4,890 English sentences with 142,116 words and 11,760 Polish sentences with 184,524 words scraped from online services via their newsfeeds over the course of five days. Sieminski points out that the complexity of the algorithm used on the Polish documents required an additional level of sophistication in terms of computing word means and disambiguation.

Traditional "manual" methods of indexing simply cannot now cope with the vast quantities of information generated on a daily basis by humanity as a whole in scientific research more specifically. The new algorithm once optimised could radically change the way in which we make archived documents searchable and allow knowledge to be extracted far more readily than is possible with standard indexing and search tools.

The approach also circumvents three critical problems faced by most users of conventional search engines: First, the lack of familiarity with the advanced search options of search engines, with a semantic algorithm advanced options become almost unnecessary. Secondly, the rigid nature of the options that are unable to catch the subtle nuance of user information needs, again a tool that understands the meaning of a search and the meaning of the results it offers avoids this problem. Finally, the unwillingness or unacceptably long time necessary to type a long query, semantically aware search will require only simply input.

Sieminski points out that the key virtue of the research is the idea of using the statistical similarity measures to assess semantic similarity. He explains that semantic similarity of words could be inferred from the WordNet database. He proposes using this database only during text indexing. "Indexing is done only once so the inevitably long processing time is not an issue," he says. "From that point on we use only statistical algorithms, which are fast and high performance."

More information: "Fast algorithm for assessing semantic similarity of texts" in *Int. J. Intelligent Information and Database Systems*, 2012, 6, 495-512

Provided by Inderscience Publishers

Citation: Fast algorithm extracts, compares document meaning (2012, September 25) retrieved 27 April 2024 from https://phys.org/news/2012-09-fast-algorithm-document.html