Ian Witten (University of Waikato)

Semantic document processing using Wikipedia as a knowledge base

Wikipedia is a goldmine of information; not just for its many readers, but also for the growing community of researchers who recognize it as a resource of exceptional scale and utility. It represents a vast investment of manual effort and judgment: a huge, constantly evolving tapestry of concepts and relations that is being applied to a host of tasks.

This talk will introduce the process of "wikification"; that is, automatically and judiciously augmenting a plain-text document with pertinent hyperlinks to Wikipedia articles -- as though the document were itself a Wikipedia article. This amounts to a new semantic representation of text in terms of the salient concepts it mentions, where "concept" is equated to "Wikipedia article." Wikification is a useful process in itself, adding value to plain text documents. More importantly, it supports new methods of document processing.

We first describe how Wikipedia can be used to determine semantic relatedness, and then introduce a new, high-performance method of wikification that exploits Wikipedia's 60 M internal hyperlinks for relational information and their anchor texts as lexical information, using simple machine learning. We discuss applications to knowledge-based information retrieval, topic indexing, document tagging, and document clustering. Some of these perform at human levels. For example, on CiteULike data, automatically extracted tags are competitive with tag sets assigned by the best human taggers, according to a measure of consistency with other human taggers.

All our work uses English, but involves no syntactic parsing, so the techniques are largely language independent. The talk will include several live demos (Deo volente). All code produced by this project is open source.