

Experiments with Semantic Relatedness Measures in WordNet and UMLS

Youbo Wang and Dr. Valerie Cross

Computer Science and Systems Analysis Department

Miami University

Oxford, OH

Many consider the Semantic Web as the next generation of the Web, one that software agents can manage in order to assist users. Ontologies have been promoted as a sound basis for communications of all types on the Web. The proliferation of multiple ontologies for heterogeneous information systems developed independently, however, requires Semantic Web technology to provide tools to enable semantic interoperability. Interoperability is established by discovering semantically appropriate mappings between different and independent ontologies. Finding this mapping requires some notion of a measure of semantic relatedness between concepts. Numerous such measures have been proposed for use within an ontology. But the majority of these measures have been based on two underlying approaches: distance-based within a network structure and information content-based on a common parent between concepts.

For the distance-based approach, the number of edges in the shortest path between the two concepts measures the distance between them. The shorter the distance the more similar the concepts are. This approach was based on a hierarchical is-a semantic network. Although this edge-counting approach is intuitive and direct, it is not sensitive to the depth of the nodes for which a distance is being calculated so numerous other measures have been proposed to overcome this limitation. The foundation for information content-based approach is the insight that conceptual similarity between two concepts may be judged by the degree to which they share information. The more information they share then the more similar they are. In an is-a hierarchical network, this common information is contained in the most specific concept that subsumes these two concepts, the lowest common subsumer. The similarity value is the information content value of the lowest common subsumer. Information content measure has been typically determined by using probability of occurrences in a corpus.

The evaluation of these measures has been done predominately through experiments using these measures on word pairs from the WordNet ontology, a large-scale lexical database of English structured as a semantic network. Recently some experiments have been developed for UMLS (Unified Medical Language System), which contains a large vocabulary database of biomedical and health-related concepts and their relationships. In these experiments, the similarity of meaning of some word pairs was rated by human judges and only the naïve edge counting measure was used.

In this work, an experimental software test bed that implements the semantic relatedness measures and automates their performance testing is proposed to evaluate the validity, performance and applicability of these measures on WordNet and UMLS. Results of such experimentation could serve as a basis for developing guidelines for using such measures in a variety of applications. Developing an experimental test bed may also aid in the creation of more application specific evaluations of semantic relatedness measures and the development of methods using these measures to assess the quality of ontologies.