

Introduction to Semantic Analysis in Modern Information Retrieval Systems

As search engines attempt to furnish increasingly relevant results for searchers, many fascinating technologies have emerged. Among the most important are those which relate to language interpretation. From what was once the inherent weakness of search indexes has emerged some of the most advanced conceptual interpretation systems that have ever existed. Furthermore the research effort in this field that has been conducted by major search engines has led to the development of language concept models that have applicability far beyond Internet search engines; providing valuable tools to scientific fields such as psychology and the social sciences.

The following is a basic overview of semantic interpretation and indexing methods commonly integrated into modern search engine architecture.

The purpose of semantic technologies is to identify and map conceptual relationships between information objects such as words or phrases to facilitate a more precise understanding of text language ontology (disambiguation process). This is accomplished by establishing a spatial model in which language information objects are arranged in hyper-dimensional space according to an established language ontology scheme. This can be visualized as a 3D map in which words are assigned specific loci; for example you would find "wood" at a specific mathematical point with the word "lumber" at a distinct location nearby. In the same area could be found "tree", however "motorcycle" would be on the other side of the semantic universe.

The disambiguation process has two complimentary applications:

Text Interpretation: Text is analyzed and if necessary reduced to core words which are then compared to a natural language ontology lexicon in an attempt to interpret the whole value concepts that are embodied by the text.

Semantic Matching: An information object such as a search query is projected into semantic space to ascertain contextual natural language interpretation.

The ontology lexicon is a massive database that understands language usage conventions and other aspects of natural language. This pre-programmed understanding is combined with probability distributions to formulate a functional ontology. Information is organized as

concepts and inter-concept relationships as well as linguistic references to assist disambiguation. The ontology is structured according to three information object components: Tokens: which correspond to individual words. Terms: which are multiple token sequences that represent stand alone units. Meanings: that define core concepts and interpretations. Each term is assigned to a meaning(s) which is represented in the database by definition and by relatedness to other terms and meanings. The types of relationships identified include: Synonymy/antonymy, Similarity, Hypernymy, Logical Membership, Metonymy, Substance Family, Product, Causation, etc. Each type of relationship between meanings is assigned a score which indicates the strength of the relationship. These strengths are used to determine semantic connectivity.

The semantic space component is particularly well suited for query interpretation and involves the reduction of a search query to tokens and then terms which are then projected into semantic space to ascertain similitude between the query and the objects comprising semantic space.