Semantic is beautiful: clustering and diversifying search results with graph-based Word Sense Induction

Roberto Navigli

Department of Computer Science, Sapienza University of Rome
navigli@di.uniroma1.it

Abstract: Web search result clustering aims to facilitate information search on the Web. Rather than presenting the results of a query as a flat list, these are grouped on the basis of their similarity and subsequently shown to the user as a list of possibly labeled clusters. Each cluster is supposed to represent a different meaning of the input query, thus taking into account the language ambiguity issue. However, Web clustering methods typically rely on some notion of textual similarity of search results. As a result, text snippets with no word in common tend to be clustered separately, even if they share the same meaning.
In this talk, we present a novel approach to Web search result clustering based on the automatic discovery of word senses from raw text, a task referred to as Word Sense Induction (WSI). Key to our approach is to first acquire the senses (i.e., meanings) of a query and then cluster the search results based on their semantic similarity to the word senses induced. Our experiments, conducted on datasets of ambiguous queries, show that our approach outperforms both Web clustering and search engines in the clustering and diversification of search results.