DD2476 Search Engines and Information Retrieval Systems

Project 3: Context Expansion

Contact: Simon Stenström, Findwise (simon.stenstrom@findwise.se, 073-616 35 34)

This project is worth 3 ECTS credits. This means that it is expected to require 80 hours of work for each person in the group. The project formulation, method, and results are presented in a report as well as in a poster session. For more details, look at the course homepage, under Project in the menu.

Problem

Free text search is great for finding text snippets in a larger text mass. Inverted indexes are fast and reliable, but have the weakness of not "understanding" the concept of what you are search for. Even though a text clearly describes what a graph is, if it doesn't contain the word "graph", it won't be returned for that query.

Assignment

Your assignment is to implement a context expander, by automatically identifying "synonyms" (or other words that often occur together with a term) to the query or index.

This can be done by extracting data from a data source (for example wikipedia) and extracting the most important concepts (for example all long word or the words with the highest tf-idf). If the concepts occur together often, they should be considered "synonym"

This assignment can also be expanded into indexing a part of Wikipedia into Apache Solr[1] and use your synonyms to improve the search experience.

About Findwise

Findwise is a growing IT consultancy company, founded in 2005 by a team of experts from the enterprise search industry. The company currently employs about 90 people (January 2012) and have offices in Sweden, Denmark, Norway and Poland.

The project is meant to be fun but could possibly be used as a demo of what our customers could do with their data more than just making it searchable. There are some

[1] http://lucene.apache.org/solr/

ideas on how to solve the problem in the text, but other solutions are also warmly welcome.

If you have any questions, don’t hesitate to ask (in Swedish or English).