DD2476 Search Engines and Information Retrieval Systems

Project 1: Image Search

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

Trying to find an image of something by describing it with words has for a long time been a hard problem to solve. A search engine would need a description to index to be able to provide any information on what the image contains. But that description is very time consuming to create.

By indexing the text surrounding an image we can hopefully get some valuable information back. When you read a news site, there are often text snippets describing the image directly after the image and on pages with a large text content there are often related images that explain the text or just make it more readable. This text content is easily indexed and should be able to provide enough information to make a lightweight Google image search.

Assignment

Your assignment would be to parse a number of pages that contains images into "documents" containing the url to the image and some surrounding text to that image. Index the document into Apache Solr[1] (for example by using the Solr Java API, SolrJ) to index the information. Create a simple application (or web application) that gives the user the opportunity to search for a word and get images back as results.

This assignment can be expanded by trying to figure out the important concepts in the text and remove all other text.

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.

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

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 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).