Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam

Informationssysteme
Prof. Dr. Felix Naumann
Masterprojekt WS 2011/12

LuSim: Similarity Search with Apache Lucene

Background

Given a set of objects, similarity search aims to find all objects similar to a given query object (1). For example, a query to a customer database may contain typos or missing values; despite these differences the corresponding database entry should be found by the search application. Another example is finding similar images in an image database, as Google demonstrates with the "Similar image" function. In most cases, similarity queries need to be answered extremely fast for a satisfying user experience.

In the field of search engines and information retrieval, there is much work on finding web pages with exact keyword queries (2). Highly optimized index structures offer fast access to relevant web pages given the keywords. In our project, we will work with Apache Lucene – the most popular open-source library containing implementations of many common search engine components and techniques. Support for similarity queries is yet extremely limited.

Description

The main objective of the project is to determine whether search engine techniques can be used to support efficient similarity queries. Search engines offer very efficient data access for exact keyword queries. We aim to take advantage of these techniques to develop an *efficient* similarity search framework.

Another objective is to support *flexible* similarity queries. Similarity search requires a similarity measure on the data. Similarity indexes usually work only with a static similarity measure. Our search framework should allow users to configure the similarity measure specifically for their queries (e.g., by specifying weights of different aspects).

Tasks in this project include:

- Implementation of efficient similarity indexes with Lucene
- Implementation of support for flexible similarity queries in Lucene
- Development of a web application demonstrating the new query capabilities
- Demo description suitable to be published at a leading database or information retrieval conference

Bibliography

1. **Zezula, Pavel, et al.** *Similarity Search: The Metric Space Approach.* USA : Springer, 2005.
2. **Croft, W. Bruce and Metzler, Donald.** *Search Engines: Information Retrieval in Practice.* USA : Addison-Wesley Publishing Company, 2009.

Contact

- Prof. Dr. Felix Naumann & Dustin Lange