Zürcher Hochschule für Angewandte Wissenschaften
Lagerstrasse 41, Postfach, CH-8021 Zürich
Telefon +41 58 934 82 40, Fax +41 58 934 82 41
info-sg.zh.engineering@zhaw.ch, www.zhaw.ch

Aufgabenstellung Seminararbeit

Thema:	Evaluierung der Retrieval-Leistung von Google und Lucene am Beispiel von News-Webseiten
Studierender:	Sandro Brunner
Betreuungsperson:	Klaus Wolfertz
Ausgangslage:	Google ist heutzutage einer der am meisten verwendeten Suchmaschinen. Suchabfragen geben nicht selten mehrere Millionen Treffer in Sekundenbruchteilen zurück. Google bietet mit Hilfe von sogenannten erweiterten Such-Operatoren die Möglichkeit, Suchen auf bestimmte Webseiten zu beschränken. Dies liefert eine überschaubare Anzahl an Resultaten und sollte es ermöglichen, die Retrieval-Leistung von Google mit der einer anderen Search-Engine zu vergleichen.
Ziel der Arbeit:	Das Ziel der Arbeit ist die Beantwortung der Fragen, wie gut die Retrieval-Leistung von Google im Vergleich mit der von Lucene abschneidet sowie ob und wie man diese Leistung verbessern kann. Such-Operationen sollen sich dabei auf einen begrenzten Dokumentenbestand beschränken.
Aufgabenstellung:	1. Beschaffung von Dokumenten von Schweizer News-Webseiten mit Hilfe eines Web-Crawlers. 2. Indizierung der in Punkt 1 beschafften Dokumente mit Hilfe von Lucene. 3. Definition und Durchführung von Suchabfragen (Queries) mit Google und Lucene. 4. Vergleich der Retrieval-Leistung von Google und Lucene. 5. Evaluierung der Möglichkeiten zur Verbesserung der Retrieval-Leistung von Google und Lucene. 6. Verbesserung der Suchabfragen (Queries) aus Punkt 3 bzw. Indizierung aus Punkt 2 und erneute Durchführung der Suchabfragen mit Hilfe der in Punkt 5 gewonnen Erkenntnisse. 7. Zusammenfassung der Erkenntnisse der Retrieval-Leistung von Google und Lucene (Stärken / Schwächen, Möglichkeiten zur Verbesserung).
Erwartete Resultate:	1. Dokumente von einer oder mehreren Schweizer News-Webseiten zur Indizierung mit Lucene. 2. Vergleich der Retrieval-Leistung (Precision/Support) für ausgewählte Suchen (Queries) von Google und Lucene. 3. Auflistung der Verbesserungsmöglichkeiten durch Änderung von Queries (Google) oder der Indizierungsstrategie (Lucene). 4. Vergleich der Retrieval-Leistung (Precision/Support) nach Verbesserung der Queries / Indizierungsstrategie. 5. Überblick/Fazit über Stärken und Schwächen sowie Verbesserungsmöglichkeiten für die Search-Engines Google und Lucene.
Geplante Termine:	13.03.2013 Kick-Off 11.06.2013 Abgabe Seminararbeit 18.06.2013 Schlusspräsentation

Student Dozent

20.04.2013

Zürcher Hochschule für Angewandte Wissenschaften
Lagerstrasse 41, Postfach, CH-8021 Zürich
Telefon +41 58 934 82 40, Fax +41 58 934 82 41
info-sg.zh.engineering@zhaw.ch, www.zhaw.ch

Sandro Brunner

Klaus Wolfertz

Studiengangleiter

Reto Knaack

20.04.2013