IRG Semesterkurzbeitrag 2013

Eingesetztes System:

Zuerst haben wir uns einen Überblick der verfügbaren IRG Systeme gemacht. Aufgrund der Programmiersprache, Dokumentation und Einsatzgebietes haben wir uns für das IRG System Lucene der Apache Foundation entschieden.

Search Engine	Storage[f]	Increm. Index	Results Excerpt	Results Template	Stop words	Filetype[e]	Stemming	Fuzzy Search	Sort[d]	Ranking	Search Type[c]	Indexer Lang.[b]	License[a]
Datapark	2	■	■	■	■	1,2,3	■	■	1,2	■	2	1	4
ht://Dig	1	■	■	■	■	1,2	■	■	1	■	2	1,2	4
Indri	1	■	■	■	■	1,2,3,4	■	■	1,2	■	1,2,3	2	3
IXE	1	■	■	■	■	1,2,3	□	■	1,2	■	1,2,3	2	8
Lucene	1	■	□	□	■	1,2,4	■	■	1	■	1,2,3	3	1
MG4J	1	■	■	■	■	1,2	■	□	1	■	1,2,3	3	6
mnoGoSearch	2	■	■	■	■	1,2	■	■	1	■	2	1	4
Namazu	1	■	■	■	□	1,2	□	□	1,2	■	1,2,3	1	4
Omega	1	■	□	■	■	1,2,4,5	■	□	1	■	1,2,3	2	4
OmniFind	1	■	■	■	■	1,2,3,4,5	■	■	1	■	1,2,3	3	5
OpenFTS	2	■	□	□	■	1,2	■	■	1	■	1,2	4	4
SWISH-E	1	■	□	□	■	1,2,3	■	■	1,2	■	1,2,3	1	4
SWISH++	1	■	□	□	■	1,2	■	□	1	■	1,2,3	2	4
Terrier	1	□	□	□	■	1,2,3,4,5	■	■	1	■	1,2,3	3	7
WebGlimpse	1	■	■[g]	■[g]	□	1,2	□	■	1[e]	■	1,2,3	1	8,9
XMLSearch	1	■	□	□	■	3	□	■	3	□	1,2,3	2	8
Zettair	1	■	■	□	■	1,2	■	□	1	■	1,2,3	1	2

[a] 1:Apache,2:BSD,3:CMU,4:GPL,5:IBM,6:LGPL,7:MPL,8:Comm,9:Free
[b] 1:C, 2:C++, 3:Java, 4:Perl, 5:PHP, 6:Tcl
[c] 1:phrase, 2:boolean, 3:wild card.
[d] 1:ranking, 2:date, 3:none.
[e] 1:HTML, 2:plain text, 3:XML, 4:PDF, 5:PS.
[f] 1:file, 2:database.
[g] Commercial version only.

■ Available
□ Not Available

Motivation:

Wie wir im IRG Unterricht gesehen haben kann die Sprache auf diverse Aspekte mehrdeutig sein. So muss man bei der Suche auf Synonyme und Homonyme achten. Es gibt Begriffe welche Sprach-, Region- und Traditionsabhängig sind, so werden in verschiedenen Sprachen Metaphern verwendet, welche nicht einfach Wort für Wort übersetzt werden können.

Ein weitere Ansporn war folgende Passage aus dem Buch „Multilingual Information Retrieval von Carol Peters l Martin Braschler l Paul Clough, Vorwort“. Aufgrund dieser Aussage wollten wir prüfen ob ein Mix aus mehr Sprachen ein besseres Resultat liefert.

As described in this book, language diversity has generated an extensive variety of challenges to be solved by computer scientists. In German, for example, compound constructions are very frequent but the real concern is that the same concept might be expressed by two (or more) formulations, thus rendering it more difficult to find useful matches between search keywords and target items. The numerous grammatical cases found in Finnish grammar represent another example. Here the real problem involves irregularities imposed by vowel harmonies, which in turn generate complex morphological processing requirements. More research is required in order to promote better automatic processing of such linguistic constructions.

In erster Linie wollten wir untersuchen, ob es einen Unterschied macht, wenn wir die bestehenden Query und Collection in unterschiedlichen Sprachen suchen. Das heisst, wir haben die Query und

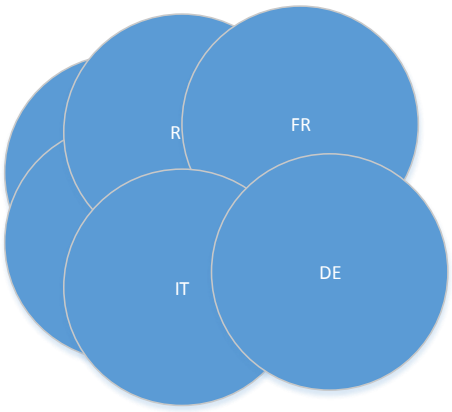
Collections genommen und übersetzt. Anschliessend haben wir untersucht, ob das Resultat der Original Query und der übersetzten Query/Collection einen Unterschied gibt.

Wenn dem nicht so ist, wäre unsere Untersuchung nicht weiter spannend und wir hätten dann etwas Neues suchen müssen. Wir haben jedoch festgestellt, dass wir mit der Suche in unterschiedlichen Sprachen auch unterschiedliche Resultate erhalten. Zusätzlich haben wir noch das Ganze mit Stopwords untersucht was ebenfalls zu anderen Resultaten führte. (Was wir hier auch erwartet hatten).

Anhand dieser Erkenntnisse haben wir uns gefragt, was für ein Resultat wir erhalten, wenn wir nun ausgewählte Sprachen verwenden und die Resultate der einzelnen Sprachen miteinander aggregieren und daraus eine neue Resultatrangliste bilden.

So haben wir entschieden, für unseren Semesterkurzbeitrag uns auf sechs verschieden Sprachen aus unterschiedlichen Sprachfamilien mit und ohne Stoppwörter zu verwenden.

Germanic(Deutsch, Englisch), Romance(French, Italian), Slavic(Russian) und Uralic(Finnish).



Die Schnittmenge der Resultate aus den einzelnen Suchen soll die neue Rangliste geben. Damit wir die Resultate zusammenführen konnten, haben wir uns entschieden die TF/IDF Scores der Resultate zusammenzuzählen. Da die Rangierung über alle Dokumente und Queries nicht aussagekräftig war.

Bsp:

DE:		Dok	Rang	Score
Query 245	Q0	124	10	1.2
FR:				
Query 245	Q0	124	200	0.1
NeueRangliste:				
Query 245	Q0	124	n	1.3

Aufbau Index:

Wir haben entschieden sechs Ranglisten der Sprachen Englisch, Deutsch, Finnisch, Französisch, Italienisch und Russisch mit- und sechs Ranglisten ohne Stoppwörter und eine Rangliste über alle Resultate zu erzeugen.

Um die Collection und Query, welche ursprünglich auf Englisch war, in die von uns gewählten Sprachen zu übersetzen mussten wir eine Lösung suchen, wie wir 20'000 Dokumente effizient und möglichst schnell übersetzen können. Übersetz wurden die Querys und Collections dann mit Hilfe von Google Translator Toolkit.

Die Liste der Stoppwörter konnten wir von http://members.unine.ch/jacques.savoy/clef/index.html entnehmen.

Interpretation der Resultate

Vergleicht man zwei Ranglisten der selben Sprache mit und ohne Stoppwörter und beobachtet dabei die 10 besten Ergebnisse, stellt man fest dass 90% der Ergebnisse bei beiden Ranglisten unter den besten 10 erscheinen. Beobachtet man nun die besten 15 Ergebnisse, stellt man fest dass nur noch 86.6% der Ergebnisse bei beiden Ranglisten unter den besten 15 enthalten sind. Analysiert man nun die Ergebnisse Stichprobenartig, stellt man fest dass die Rangliste mit Stoppwörtern die bessere ist, als ohne Stoppwörter und einen höheren Score erzielt.

Vergleicht man nun die Rangliste über alle Resultate mit den einzelnen Resultaten, stellt man fest, dass der Scor bei der Master Rangliste noch höher ist als bei den anderen.