Information Retrieval - Sommer 2006
Dipl.-Inform. Sascha Kriewel, LF 137
kriewel@is.informatik.uni-duisburg.de

Übungsblatt 6 — IR-Modelle

Aufgabe 15: Boolesches Retrieval

(a) Was sind Vor- und Nachteile des Booleschen Retrievals.

Die IR-Forschung ist schon lange zu dem Schluß gekommen, dass das Boolesche Modell recht ungeeignet für die Anwendung im Information Retrieval ist. Warum wird es trotzdem noch in vielen Anwendungen eingesetzt?

(b) Angenommen, die folgenden Dokumente seien von einem Booleschen Retrievalsystem indexiert worden. Dabei fand die übliche Stoppworteliminierung, sowie Stemming statt.

1. Evaluating Strategic Support for Information Access in the Daffodil System.
2. Daffodil: A User-Oriented Approach for Accessing Federated Digital Libraries.
3. Daffodil: Distributed Agents for User-Friendly Access of Digital Libraries.
4. Daffodil - Strategic Support for User-Oriented Access to Heterogeneous Digital Libraries.
5. Active Support for Query Formulation in Virtual Digital Libraries: A case study with Daffodil.
6. Daffodil: An integrated desktop for supporting high-level search activities in federated digtal libraries.
7. User-Oriented Query Modification in Metaclass Systems.
8. Daffodil: Strategic Support Evaluated

Formuliere nun möglichst knappe Boolesche Anfragen (mit AND, OR und NOT), die genau die folgenden Dokumente finden:

(*i*) 2 und 4

(*ii*) 8

(*iii*) 1, 3 und 6

(*iv*) 5 und 8

Aufgabe 16: Vektorraummodell

Gegeben sei ein IR-System, welches für das Retrieval von Dokumenten das Vektorraummodell benutzt. Die folgenden Dokumentrepräsentationen seien

vorhanden (Zahl hinter jedem Term gibt die Anzahl des Vorkommens an, wenn von 1 verschieden):

D1 „Europa (3) Parlament Bürger Verfassung"

D2 „Europa (2) Partei Partner Union Bürger"

D3 „Europa Volk Einheit Bürger (2)"

Das Vokabular soll aus der Vereinigung der Terme in den Dokumenten nach Stoppworteliminierung und Grundformreduktion entstanden sein:

$T =$ { Bürger, Einheit, Europa, Parlament, Partei, Partner, Union, Verfassung, Volk }

(a) Stelle zunächst die Vektoren für die Dokumentenkollektion nur unter Berücksichtigung der Termnhäufigkeit auf.

(b) Berechne dann die Vektoren für die Dokumentenkollektion nach der $tf \cdot idf$-Formel:

$w_{mi} = ntf_i \cdot idf_i$

(c) In den Anfragevektoren zu den folgenden beiden Anfragen sollen alle Terme gleichgewichtet sein.

Q1 Europa Union

Q2 Verfassung Bürger Europa

(d) Ermittle nun für beide Varianten der Dokumentenvektoren die Ähnlichkeit zwischen den Anfragen und den Dokumenten mit Hilfe des Skalarprodukts. Ergebnis dieser Aufgabe sollten jeweils sechs Ähnlichkeitswerte $sim(q_i, d_j)$ sein.

Vergleiche das Ranking der Ergebnisse. Wie kommt der Unterschied zustande? Welches der Ergebnisrankings ist für den Benutzer wohl zufriedenstellender?

Tip: Man kann sich viel stupide Rechenarbeit sparen, wenn man diese Aufgabe mit einer Tabellenkalkulation löst.

Aufgabe 17: Ein einfaches IR-System mit Lucene
Abgabe bis 26. Mai 2006, 12 Uhr:
`iruebg-abgaben@is.informatik.uni-duisburg.de`

Erweitere Deine Lösung aus Aufgabe 14 (Blatt 5) um die folgenden Aspekte.

- Die Bestimmung der gestemmten Wortlisten mit Häufigkeiten soll nun nicht nur für einen Artikel, sondern für mehrere Dokumente durchgeführt werden. Die Dokumente (mindestens 10, z.B. Artikel aus der deutschsprachigen Wikipedia) dürfen fest vorgegeben sein.
- Für eine einfache Boolesche Anfrage mit maximal zwei Termen (a, $a \wedge b$, $a \vee b$, $a \wedge \neg b$, $a \vee \neg b$, ...) sollen die passenden Dokumente aus den indexierten Artikeln gefunden werden.

- *Coordination Level Match* ist ein vereinfachter Spezialfall des Vektorraum-Modells. Setze es um, so dass zu einer Liste von Anfragetermen eine gerankte Liste der indexierten Dokumente zurückgeliefert wird.

Beachte, dass auch für die Terme der Anfrage Stoppworteliminierung und Stemming durchgeführt werden müssen.