**Übungen zu Internet-Suchmaschinen, Sommersemester 2015**

Vu Tran (LF 139)
Sprechstunde nach Vereinbarung
vtran@is.inf.uni-due.de

# Übungsblatt 4

**keine Abgabe**

## Aufgabe 9: 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. Erläutere in drei bis fünf Sätzen, warum es trotzdem noch in vielen Anwendungen eingesetzt wird.

(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 Desktop 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: Integrated Search Support Evaluated.

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

(*i*) 4

(*ii*) 6, 7 und 8

(*iii*) 2 und 3

### Aufgabe 10: Vektorraummodell

Gegeben sei ein IR-System, das für das Retrieval von Dokumenten das Vektorraummodell benutzt. Die folgenden Dokumentrepräsentationen seien vorhanden. Die Terme sind durch Kommata getrennt und die Zahl hinter jedem Term gibt die Anzahl des Vorkommens an, wenn von 1 verschieden:

D1 „retrieval (20), digital libraries, interface (6), evaluation (2)"

D2 „digital libraries (3), evaluation (2), retrieval (17), interface, user, service"

D3 „agents, access, retrieval (22), distributed"

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

$T =$ { access, agent, digital library, distributed, evaluation, interface, retrieval, service, user }

(a) Stelle zunächst die Vektoren für die Dokumentenkollektion nur unter Berücksichtigung der Termhäufigkeit 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 der folgenden Anfrage sollen alle Terme gleichgewichtet sein.

Q1 digital libraries, retrieval

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

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

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