Datenbanken & Informationssysteme
Übungen Teil 4

Information Retrieval

[Einige der folgenden Aufgaben stammen von Norbert Fuhr, Uni Duisburg-Essen]

1. Anwendungen des Information Retrievals

Beispiele für Systeme des Information Retrieval sind:

- Web-Suchmaschinen
- Suche in Online-Dokumentationen
- Digitale Bibliotheken
- Suche in Bildarchiven

Finden Sie jeweils ein Beispiel und untersuchen Sie

- Welche Dokumente oder Informationsobjekte können in dem jeweiligen System gefunden werden?
- Nach welchen Kriterien können die Objekte gesucht werden?
- Wie sieht das Ergebnis einer Anfrage aus?
- Wie beurteilen Sie die Antworten in Bezug auf Effizienz und Effektivität?

2. Vektorraummodell

Gegeben seien folgende Dokumentrepräsentationen (entstanden durch Extraktion aus den drei Original-Dokumenten). Dabei gibt die Zahl beim Term an, wie oft er im Dokument vorkommt.

D_1 „retrieval, digital libraries, interface (3), evaluation“

D_2 „evaluation (2), retrieval, interface, user, service“

D_3 „digital libraries (3), agents, access, retrieval (2), distributed“

Die Terme des Vokabulars sind:

> „access, agents, digital libraries, distributed, evaluation, interface, retrieval, service, user“

Aufgabe:

(a) Bestimmen Sie N_t (*document frequency*), die Zahl der Dokumente, die Term t enthalten.

Lösung

Bezogen auf das obige Vokabular in alphabetischer Reihenfolge:

```
( 1 1 2 1 2 2 3 1 1 )
```

(b) Ermitteln Sie die Vektoren für die Dokumente (mit Berücksichtigung der Termhäufigkeit)

Lösung

$\vec{D_1} = (\ 0\ 0\ 1\ 0\ 1\ 3\ 1\ 0\ 0)$
$\vec{D_2} = (\ 0\ 0\ 0\ 0\ 2\ 1\ 1\ 1\ 1)$
$\vec{D_3} = (\ 1\ 1\ 3\ 1\ 0\ 0\ 2\ 0\ 0)$

(c) Betrachten Sie folgende Anfragen:

- Q_1 „retrieval, evaluation“
- Q_2 „digital libraries, interface, evaluation“

Berechnen Sie die „Ähnlichkeit“ zwischen diesen Anfragen und den Dokumenten.

(d) Interpretieren Sie die Ergebnisse.

3. Invertierter Index

Bilden Sie einen invertierten Index mit dem Aufbau

Term: $\langle$ Dok$_1$: Pos$_1$, Pos$_2$, ...; Dok$_2$: Pos$_1$, ... $\rangle$

aus folgenden Dokumenten:

1: Frankreich ist ein europäisches Land

2: Deutschland und Frankreich sind benachbart

3: Paris ist die Hauptstadt von Frankreich

4: Paris ist eine Weltstadt, Frankfurt auch

Als Liste der *stop words* sei „ist, ein, und, sind, die, von, eine, auch“ vorgegeben.

4. Suchen im invertierten Index

Gegeben sei ein invertierter Index mit dem Aufbau

Term: $\langle$ Dok$_1$: Pos$_1$, Pos$_2$, ...; Dok$_2$: Pos$_1$, ... $\rangle$,

nämlich:

was:	$\langle$ 1: 1,14,102,302; 3: 11,53,233,401; 6: 26,43,82; 8: 23,49,401 $\rangle$
du:	$\langle$ 2: 63,105,282; 3: 12,88,143; 6:27,128,169,482 $\rangle$
heute:	$\langle$ 1: 211,234,311; 3: 13; 4: 100,122; 6: 28,234 $\rangle$
kannst:	$\langle$ 2: 179,284; 3: 14,87,156; 6: 29,70 $\rangle$
besorgen:	$\langle$ 3: 15; 6: 30,155; 7:67,166 $\rangle$
verschiebe:	$\langle$ 3: 17,53; 5: 40,99,120; 8: 45,132 $\rangle$
nicht:	$\langle$ 3: 18,44,217; 4: 34,97; 8: 1,46,156 $\rangle$
auf:	$\langle$ 3: 19,61,101,189; 8: 47,386 $\rangle$
morgen:	$\langle$ 3: 20,111,273; 4: 24,103; 8:48,430 $\rangle$

Ermitteln Sie die Fundstellen für folgende Anfragen:

(a) besorgen

(b) besorgen and was

(c) „verschiebe nicht auf morgen“ (als Phrase)

(d) „was du heute kannst“

(e) „heute kannst besorgen“ and „verschiebe nicht auf morgen“