ETH Eidgenössische Technische Hochschule Zürich
Ecole polytechnique fédérale de Zurich
Politecnico federale di Zurigo
Swiss Federal Institute of Technology Zurich

Prof. P. Schäuble *A. Feßler, E. Munteanu, U. Röhm, H. Schuldt, R. Weber*

Informationssysteme IS-G SS 1997

Übung 6 Abgabe: 20.5.1997

Aufgabe 1: Ausbeute und Präzision

a) Es seien $C_1 > 0$ die Kosten, welche entstehen, wenn ein Benutzer ein vom Suchsystem gefundenes Dokument inspiziert, das irrelevant ist. Ausserdem seien $C_2 > 0$ die Kosten, die dadurch entstehen, wenn das Suchsystem ein relevantes Dokument nicht unter den ersten r Dokumenten rangiert, die der Benutzer inspiziert. Ermitteln Sie, ob in folgenden Situationen gilt $C_1 \ll C_2$, $C_1 \approx C_2$ oder $C_1 \gg C_2$.

- Ein Patentanwalt überprüft, ob eine Erfindung bereits patentiert ist.
- Ein Journalist möchte herausfinden, wann die Swissair das erste Mal über den Atlantik flog.

b) Eine Dokumentenkollektion enthalte 1 Mio. Dokumente, wovon der Benutzer 10 inspiziert. Dabei wird eine Ausbeute von 20% und eine Präzision von 50% erzielt. Wie hoch ist der Fallout (Anteil irrelevanter Dokumente, die inspiziert wurden)? Wie verhält sich der Fallout für $r \to n$?

Aufgabe 2: ff*idf - Gewichtung

Gegeben sei eine Anfrage q und $n = 4$ Dokumente d_0, d_1, d_2 und d_3 mit folgenden Token (Vorkommen eines Indexierungsmerkmals):

$$\begin{aligned} q &= < \varphi_0, \varphi_1, \varphi_2 > \\ d_0 &= < \varphi_g, \varphi_2, \varphi_2 > \\ d_1 &= < \varphi_h, \varphi_1, \varphi_2 > \\ d_2 &= < \varphi_0, \varphi_0, \varphi_0, \varphi_0, \varphi_0, \varphi_1, \varphi_1 > \\ d_3 &= < \varphi_i, \varphi_1 > \end{aligned}$$

Bestimmen Sie g, h und i, so dass bei einer ff*idf - Gewichtung, die Dokumente d_0, d_1, d_2 und d_3 mit dem Skalarprodukt auf die Ränge 1, 2, 3 und 4 zu liegen kommen.

Aufgabe 3: Antwortzeit

Berechnen Sie die erwartete Antwortzeit unter folgenden Annahmen:

- Die Anfrage q besteht aus k Indexierungsmerkmalen.
- Für die Bearbeitung eines Indexierungsmerkmals φ_i wird die Zeit $df(\varphi_i) \cdot \Delta$t benötigt.
- Dokumentenhäufigkeiten $df(\varphi_i) = \frac{c}{i+1}$ sind wie Kollektionshäufigkeiten verteilt (Zipfsches Gesetz).

a) Berechnen Sie die erwartete Antwortzeit unter der Annahme, dass die m Indexierungsmerkmale gleichwahrscheinlich als Anfragemerkmale gewählt werden: $P(\varphi_i) = \frac{1}{m}$

b) Wie verhält sich die erwartete Antwortzeit für $m \to \infty$?