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 Beispiellösung

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.

Im ersten Fall (Patentanwalt) gilt $C_1 \ll C_2$. Die Kosten C_1 (Kosten für das Inspizieren eines irrelevanten Dokuments) entsprechen dabei den Kosten für etwa 5 sec Arbeitszeit. Falls jedoch ein relevantes Dokument übersehen wird, können sich die Kosten C_2 (Kosten, die durch das Nicht-Inspizieren eines relevanten Dokuments entstehen) hingegen auf mehrere Tausend Franken (Gerichtskosten, etc.) belaufen.
Im Falle der journalistischen Recherche gilt $C_1 \gg C_2$. C_1 entspricht dabei wieder den Kosten für etwa 5 sec Arbeitszeit. Allerdings ist es in diesem Beispiel egal, ob noch ein zusätzliches relevantes Dokument existiert, das nicht inspiziert wird. Daher gilt $C_2 = 0$.
Im ersten Beispiel (Patentanwalt) ist also eine gute Ausbeute (möglichst **alle** relevanten Dokumente) gefragt, während im zweiten Beispiel (journalistische Recherche) eine gute Präzision (möglichst **keine** irrelevanten Dokumente) vorliegen sollte.

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 inspizierter irrelevanter Dokumente bzgl. aller irrelevanten)? Wie verhält sich der Fallout für $r \to n$?

Bei einer Präzision $\pi = 0.5$ gilt, dass von 10 inspizierten Dokumenten 5 relevant sind. Die Ausbeute $\rho = 0.2$ (Anteil inspizierter relevanter Dokumente bzgl. aller relevanten) gibt an, dass es insgesamt 25 relevante Dokumente gibt.

Im folgenden bezeichne n die Gesamtzahl der Dokumente, r die Anzahl der inspizierten Dokumente, $|D_r^{rel}|$ die Anzahl der relevanten Dokumente, die inspiziert wurden und $|D^{rel}|$ die Gesamtzahl der relevanten Dokumente.
Dann gilt für den Fallout: $Fallout = \frac{r - |D_r^{rel}|}{n - |D^{rel}|}$, in unserem Beispiel also:
$Fallout = \frac{10-5}{10^6-25}$
Damit ergibt sich für $r \to n$ folgendes Verhalten des Fallouts:

$$\lim_{r \to n} Fallout = \lim_{r \to n} \frac{r - |D_r^{rel}|}{n - |D^{rel}|} = \frac{10^6 - 25}{10^6 - 25} = 1$$

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.

Es sei $\varphi_g = \varphi_h = \varphi_i = \varphi_0$. Dann gilt für die Merkmalshäufigkeiten $ff(\varphi_j, d_k)$, die Dokumentenhäufigkeiten $df(\varphi_j)$ sowie die inversen Dokumentenhäufigkeiten $idf(\varphi_j) = \log(\frac{n+1}{df(\varphi_j)+1})$:

	φ_0	φ_1	φ_2
$ff(\varphi_j, d_0)$	1	0	2
$ff(\varphi_j, d_1)$	1	1	1
$ff(\varphi_j, d_2)$	5	2	0
$ff(\varphi_j, d_3)$	1	1	0
$ff(\varphi_j, q)$	1	1	1
$df(\varphi_j)$	4	3	2
$idf(\varphi_j)$	$\log(\frac{5}{5}) = 0$	$\log(\frac{5}{4})$	$\log(\frac{5}{3})$

mit den inversen Dokumentenhäufigkeiten $idf(\varphi_2) > idf(\varphi_1) > idf(\varphi_0)$

Damit gilt für die einzelnen Retrieval-Statuswerte:

$$RSV(q, d_0) = (2 \cdot idf(\varphi_2)) \cdot (1 \cdot idf(\varphi_2)) = 2 \cdot \log^2(\frac{5}{3})$$

$$RSV(q,d_1) = (idf(\varphi_1)) \cdot (idf(\varphi_1)) + (idf(\varphi_2)) \cdot (idf(\varphi_2)) = \log^2(\frac{5}{4}) + \log^2(\frac{5}{3})$$

$$RSV(q,d_2) = (2 \cdot idf(\varphi_1)) \cdot (idf(\varphi_1)) = 2 \cdot \log^2(\frac{5}{4})$$

$$RSV(q,d_3) = (idf(\varphi_1)) \cdot (idf(\varphi_1)) = \log^2(\frac{5}{4})$$

Da $RSV(q,d_0) > RSV(q,d_1) > RSV(q,d_2) > RSV(q,d_3)$ gilt, wird die gesuchte Rangierung erreicht.
Das Merkmal φ_0 ist in allen Dokumenten enthalten und trägt damit nicht zu einer aussagekräftigen Indexierung bei. Daher ergibt sich der zugehörige idf-Wert, der die Qualität des Anfragemerkmals angibt, zu Null ($idf(\varphi_0) = 0$).
Beachtenswert ist hier noch die Tatsache, dass in obigem Beispiel das Dokument d_0 vor d_1 rangiert wird, obwohl d_1 alle drei Anfragemerkmale anthält, d_0 jedoch nur zwei davon.

Aufgabe 3: Antwortzeit

Berechnen Sie die erwartete Antwortzeit T für eine Anfrage q unter folgenden Annahmen:

- Die Anfrage q besteht aus k Anfragemerkmalen aus den insgesamt m 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), wobei c eine Konstante ist.

a) Berechnen Sie die erwartete Antwortzeit unter der Annahme, dass die m Indexierungsmerkmale gleich wahrscheinlich als Anfragemerkmale gewählt werden: $P(\varphi_i) = \frac{1}{m}$
Für den Erwartungswert $E[T]$ der Antwortzeit gilt im Fall $k = 1$:

$$E[T] = \sum_t (t \cdot P(T=t)) = \sum_{i=0}^{m-1} (df(\varphi_i)\Delta t \cdot \frac{1}{m}) = \sum_{i=0}^{m-1} \frac{c}{i+1}\Delta t \cdot \frac{1}{m}$$

Die Auswahl der Anfragemerkmale erfolge unabhängig voneinander, d.h. gleiche Indexierungsmerkmale können mehrmals in der Anfrage vorkommen. Da außerdem nach Voraussetzung alle m Indexierungsmerkmale gleich wahrscheinlich sind, gilt für den Erwartungswert der Antwortzeit von Anfrage q:
$E[T] = k \cdot \sum_t (t \cdot P(T=t))$.
Mit $T = \frac{c}{i+1}\Delta t$ und der Wahrscheinlichkeit $P(\varphi_i) = \frac{1}{m}$ gilt daher:

$$E[T] = k \cdot \sum_{i=0}^{m-1} \frac{c}{i+1}\Delta t \cdot \frac{1}{m} = \frac{k \cdot c \cdot \Delta t}{m} \sum_{i=0}^{m-1} \frac{1}{i+1} = \frac{k \cdot c \cdot \Delta t}{m} \sum_{i=1}^{m} \frac{1}{i} \approx k \cdot c \cdot \Delta t \frac{\gamma + \ln(m)}{m}$$

mit der Abschätzung $\sum_{i=1}^{m} \frac{1}{i} \approx \gamma + \ln(m)$ für grosse m (γ ist die Eulersche Konstante).

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

$$\lim_{m\to\infty} E[T] = \lim_{m\to\infty} k \cdot c \cdot \Delta t \frac{\gamma + \ln(m)}{m} = 0 \quad \left(da\ \frac{ln(x)'}{x'} = \frac{1}{x}\right)$$