Max-Planck-Institut Saarbrücken
AG5: Datenbanken und Informationssysteme
Dozenten: Prof. Dr.-Ing. G. Weikum, Dr. Ralf Schenkel
Betreuer: Dipl.-Inform. Sergej Sizov, Dipl.-Inform. Christian Zimmer

Informationssysteme (SS 05)
Beispiellösungen zu Übungsblatt 1

26.April 2005

Aufgabe 1.1: Relevanzbewertung

a) Beide Ansätze berechnen einen Durchschnitt, allerdings sind diese unterschiedlich gewichtet. Während beim Makrodurchschnitnitt die Ergebnislänge keine Rolle spielt, so werden beim Mikrodurchschnitt längere Ergebnislisten deutlich stärker gewertet. Verdeutlicht wird dies durch ein Beispiel: Zwei Anfragen q_1 und q_2 liefern zwei Ergebnislisten. Bei Anfrage q_1 sind von 2 Treffern 1 relevanter Treffer vorhanden und bei Anfrage q_2 sind von 1000 Treffern sogar 999 relevant. Als Ergebnis erhält man einen deutlich höhere Mikrodurchschnitt, da dort die zweite Anfrage höher gewichtet wird und die Realität dadurch besser widergespiegelt wird.

b) Die Mikrobewertung eines IR-Systems ist hinsichtlich einer Anfragemenge nicht monoton, d.h. das Ranking der Güte zweier IR-Systeme kann sich ändern durch Hinzufügen einer Anfrage mit identischem Resultat für beide Systeme.

Beweis: Der Beweis ist vollbracht, falls ein Beispiel gefunden wird, bei dem das Hinzufügen einer Anfrage mit identischem Resultat für beide Systeme die Reihenfolge der beiden Gütewerte vertauscht. Für eine Anfragemenge $\{q_1, \ldots, q_n\}$ ist die

$$\text{Mikrobewertung} = \frac{\sum_{i=1}^{n} \text{Anzahl der für } q_i \text{ relevante und gefundene Dokumente}}{\sum_{i=1}^{n} \text{Anzahl aller für } q_i \text{ gefundenen Dokumente}} \tag{1}$$

Seien IR_1 und IR_2 zwei IR-Systeme und $\{q_1, \ldots, q_n\}$ die Anfragemenge. Seien die Summe der relevanten und gefundenen Dokumente von IR_1 für $\{q_1, \ldots, q_n\}$ gleich 10 und die von IR_2 gleich 1 sowie die Summe der gefundenen Dokumente von IR_1 gleich 15 und die von IR_2 gleich 2.

Damit ist für die Anfragemenge $\{q_1, \ldots, q_n\}$ die Mikrobewertung für IR_1 gleich $\frac{2}{3} = 0,\overline{6}$ und für IR_2 gleich $\frac{1}{2} = 0,5$.

Nimmt man nun eine Anfrage q_{n+1} hinzu, bei der für beide IR-Systeme die Anzahl der relevanten und gefundenen Dokumente 10 und die Anzahl der gefundenen Dokumente ebenfalls 10 ist, so ist die Mikrobewertung der Anfragemenge $\{q_1, \ldots, q_{n+1}\}$ von IR_1 gleich $\frac{4}{5} = 0,8$ und die von IR_2 gleich $\frac{11}{12} = 0,91\overline{6}$ und damit größer.

Aufgabe 1.2: Ähnlichkeitsmaße

Zu zeigen: Für normalisierte Vektoren ergeben das Cosinus-Ähnlichkeitsmaß und die Euklidische Distanz dasselbe Ranking. Die Euklidische Distanz zweier Vektoren $\vec{v}_1$ und $\vec{v}_2$ ist

$$|\vec{v}_1 - \vec{v}_2| = \sqrt{(\vec{v}_1 - \vec{v}_2)^2} \tag{2}$$

Das Cosinus-Ähnlichkeitsmaß zweier normierter Vektoren $\vec{v}_1$ und $\vec{v}_2$ ist

$$\begin{aligned} \cos\alpha &= \frac{\vec{v}_1 \cdot \vec{v}_2}{\underbrace{|\vec{v}_1|}_{1} \cdot \underbrace{|\vec{v}_1|}_{1}} \\ &= \vec{v}_1 \cdot \vec{v}_2 \end{aligned} \tag{3}$$

Da der Cosinus zweier identischer Vektoren 1 ist und die Euklidische Distanz aber 0 ist, d.h. es muss folgender Zusammenhang hergeleitet werden:

$$\begin{array}{rcll} \sqrt{(\vec{v}_1 - \vec{q})^2} & > & \sqrt{(\vec{v}_2 - \vec{q})^2} & \Leftrightarrow \quad \vec{v}_1 \cdot \vec{q} < \vec{v}_2 \cdot \vec{q} \\ (\vec{v}_1 - \vec{q})^2 & > & (\vec{v}_2 - \vec{q})^2 & \Leftrightarrow \quad \vec{v}_1 \cdot \vec{q} < \vec{v}_2 \cdot \vec{q} \\ \sum_{i=1}^{n} \left(\vec{v}_{1,i}^2 - 2\vec{v}_{1,i} \cdot \vec{q}_i + \vec{q}_i^2\right) & > & \sum_{i=1}^{n} \left(\vec{v}_{2,i}^2 - 2\vec{v}_{2,i} \cdot \vec{q}_i + \vec{q}_i^2\right) & \Leftrightarrow \quad \vec{v}_1 \cdot \vec{q} < \vec{v}_2 \cdot \vec{q} \\ 1 - 2\vec{v}_1 \cdot \vec{q} + 1 & > & 1 - 2\vec{v}_2 \cdot \vec{q} + 1 & \Leftrightarrow \quad \vec{v}_1 \cdot \vec{q} < \vec{v}_2 \cdot \vec{q} \\ -2\vec{v}_1 \cdot \vec{q} & > & -2\vec{v}_2 \cdot \vec{q} & \Leftrightarrow \quad \vec{v}_1 \cdot \vec{q} < \vec{v}_2 \cdot \vec{q} \\ \vec{v}_1 \cdot \vec{q} & < & \vec{v}_2 \cdot \vec{q} & \Leftrightarrow \quad \vec{v}_1 \cdot \vec{q} < \vec{v}_2 \cdot \vec{q} \end{array} \tag{4}$$

Aufgabe 1.3: Vektorraummodell

Gegeben: Ein Korpus bestehend aus den Dokumenten d_1 bis d_4:

d_1: Marcus tried to assissinate Caesar.

d_2: Marcus was a Roman.

d_3: Caesar was a ruler. All Romans were either loyal to Caesar or hated him.

d_4: Everyone is loyal to someone. People only try to assissinate rulers they are not loyal to.

Weiter sind für die Teilaufgabe c) zwei Anfragen

q_1: Who assissinated Caesar?

q_2: Loyality and assassination.

gegeben.

Nach Entfernen der Stoppwörter und der Reduktion auf die entsprechenden Wortstämme lassen sich durch die Formel

$$idf_i = \frac{N}{df_i} \tag{5}$$

die *idf*-Werte (Teilaufgabe a) aller Terme berechnen. Weiter werden mit den folgenden Formeln die Werte tf_i (normalisiert) und idf_i (gedämpft) sowie die gewichteten Dokumentenvektoren berechnet (Teilaufgabe b).

$$tf_{ij,\,\text{normalisiert}} = \frac{tf_{ij}}{\max_k tf_{kj}} \tag{6}$$

$$idf_{i,\,\text{gedämpft}} = log_2 \frac{N}{df_i} \tag{7}$$

$$w_{ij} = tf_{ij,\,\text{normalisiert}} \cdot idf_{i,\,\text{gedämpft}} \tag{8}$$

Die entsprechenden Werte sind in den beiden folgenden Tabellen dargestellt:

Term	i	df_i	idf_i		absolute tf_{ij}				normalisierte tf_{ij}			
			absolut	gedämpft	tf_{i1}	tf_{i2}	tf_{i3}	tf_{i4}	tf_{i1}	tf_{i2}	tf_{i3}	tf_{i4}
Marcus	1	2	2	1	1	1	0	0	1	1	0	0
try	2	2	2	1	1	0	0	1	1	0	0	$\frac{1}{2}$
assassin	3	2	2	1	1	0	0	1	1	0	0	$\frac{1}{2}$
Caesar	4	2	2	1	1	0	2	0	1	0	1	0
Rome	5	2	2	1	0	1	1	0	0	1	$\frac{1}{2}$	0
rule	6	2	2	1	0	0	1	1	0	0	$\frac{1}{2}$	$\frac{1}{2}$
loyal	7	2	2	1	0	0	1	2	0	0	$\frac{1}{2}$	1
hate	8	1	4	2	0	0	1	0	0	0	$\frac{1}{2}$	0
people	9	1	4	2	0	0	0	1	0	0	0	$\frac{1}{2}$

Term	i	w_{i1}	w_{i2}	w_{i3}	w_{i4}	q_{i1}	q_{i2}
Marcus	1	1	1	0	0	0	0
try	2	1	0	0	$\frac{1}{2}$	0	0
assassin	3	1	0	0	$\frac{1}{2}$	1	1
Caesar	4	1	0	1	0	1	0
Rome	5	0	1	$\frac{1}{2}$	0	0	0
rule	6	0	0	$\frac{1}{2}$	$\frac{1}{2}$	0	0
loyal	7	0	0	$\frac{1}{2}$	1	0	1
hate	8	0	0	1	0	0	0
people	9	0	0	0	1	0	0

Wird die Cosinus-Ähnlichkeit als Ähnlichkeitsfunktion verwendet, so ist

$$\text{sim}\left(\vec{d}, \vec{q}\right) = \frac{\vec{d} \cdot \vec{q}}{\sqrt{\vec{d}^2 \cdot \vec{q}^2}}. \tag{9}$$

Die sich daraus ergebenden Ähnlichkeiten sind in der folgenden Tabelle aufgeführt.

$\text{sim}\left(\vec{d}_i, \vec{q}_j\right)$	q_1	q_2
d_1	$\frac{1}{\sqrt{2}} \approx 70\%$	$\frac{1}{2\sqrt{2}} \approx 35\%$
d_2	0	0
d_3	$\frac{2}{\sqrt{22}} \approx 43\%$	$\frac{1}{\sqrt{22}} \approx 21\%$
d_4	$\frac{1}{\sqrt{22}} \approx 21\%$	$\frac{3}{\sqrt{22}} \approx 64\%$

Damit ergeben sich folgende Ranglisten für q_1 und q_2:

q_1: d_1, d_3, d_4

q_1: d_4, d_1, d_3

Dokument d_2 ist für keine der beiden Anfragen relevant und wird deswegen nicht in die Ranglisten mitaufgenommen.

Aufgabe 1.4: Threshold Algorithmus

Wir zeigen dies durch einen Widerspruchsbeweis, indem wir annehmen, dass der Algorithmus gestoppt hat und die *top-k* Dokumente festgelegt hat, es aber ein Dokument d_x gibt, welches bislang nicht in die Menge der *top-k* Dokumente aufgenommen wurde, aber eigentlich zu dieser Menge gehört. Sei d_{topk} ein beliebiges Dokument innerhalb der *top-k* Dokumente und dann gilt, da der Algorithmus abgebrochen wurde:

$$\geq worstscore(d_{topk}) \geq bestscore(d_x) \tag{10}$$

Daraus folgt dann:

$$agg_i s_i(q, d_{topk}) \geq agg_{i \in E(d_{topk})} s_i(q, d_{topk}) \geq agg_{i \in E(d_x)} s_i(q, d_x) + agg_{i \notin E(d_x)} high_i \geq agg_i s_i(q, d_x) \quad (11)$$

Somit haben wir den Widerspruch, dass Dokument d_x doch nicht in die Menge der *top-k* Dokumente gehört und bewiesen, dass der Algorithmus nicht zu frh terminiert und somit korrekt arbeitet.

Aufgabe 1.5: Erweiterter Threshold Algorithmus für Web-Portale

a) Der Threshold Algorithmus aus der Vorlesung (Folie 2-22) kann durch die Unterscheidung von R-Sources und SR-Sourcen abgeändert werden. Die Idee dabei liegt darin, die *random accesses* auf den R- und SR-Sources durchzuführen und das parallele Scannen der Sources auf die SR-Sources zu beschränken. Der entscheidende Unterschied liegt in der Initialisierung der maximalen Werte einer Source. Während bei den SR-Sourcen durch das sortierte Scannen eine Absenkung des maximalen Wertes $high_i$ möglich ist, muss für die R-Sourcen der Wert von $high_i$ konstant auf dem maximalen Wert bleiben.

b) Mögliche Ansatzpunkte: Eine SR-Source bietet sowohl für den sortierten Zugriff als auch für den wahlfreien (random) Zugriff Möglichkeiten. Der random access sollte immer dann ausgenutzt werden wenn er möglich ist. Dagegen bieten R-Sources nur den random access an, so dass die Annahme, dass alle folgenden Werte kleiner sind, nicht gemacht werden kann.

c) Eigentlich kann man den allgemeinen Treshold Algorithmus auch durch Hinzunahme der S-Sources anwenden. Aber es bietet sich an, die Variante mit sortiertem Zugriff zu verwenden (Folie 2-23). Hierbei erfolgt das parallele Scannen über alle S- und SR-Sources und die Initialisierung der $high_i$ Werte der R-Sources muss wieder mit dem maximalen Wert erfolgen, welcher im Verlauf des Algorithmus nicht angepasst wird.

Aufgabe 1.6: LSI

a) $q = d_4^T = (1\ 0\ 0\ 0\ 2\ 0\ 0\ 1\ 0\ 0\ 0\ 0)^T$

$$\begin{aligned} q' &= q \times U_3 = & (12)\\ &= (0.22 + 1.28 + 0.3; -0.11 - 0.34 - 0.14; 0.29 + 0.72 + 0.33) = & (13)\\ &= (1.8; -0.6; 1.3) & (14)\\ & & (15) \end{aligned}$$

Wir benutzen das Skalarprodukt, um die Ähnlichkeit zu berechnen:

$$\begin{aligned} sim(q', d_1') &= 0,5 & (16)\\ sim(q', d_2') &= 0,3 & (17)\\ sim(q', d_3') &= 1,2 & (18)\\ sim(q', d_4') &= 1,8 & (19)\\ sim(q', d_5') &= -0,2 & (20)\\ sim(q', d_6') &= 0,0 & (21)\\ sim(q', d_7') &= 0,0 & (22)\\ sim(q', d_8') &= 0,0 & (23)\\ sim(q', d_9') &= 0,0 & (24) \end{aligned}$$

b) $q = (1\ 1\ 0\ 1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0)^T$

	$q' \approx (0.8; -0.1; \mathbf{0})$	$q' \approx (0.8; -0.1; \mathbf{0.1})$
d_2	0.47	0.42
d_3	0.38	0.40
d_4	0.45	0.51

Dokumente mit einem höheren Wert bei den ersten Komponmenten liefern offensichtlich ein besseres Ranking: d_2 and d_4. Eine ausführliche Berechnung der Ähnlichkeiten ist nicht notwendig.

Aufgabe 1.7: LSI

$$
\begin{aligned}
A_2 &= U_2 \times \Delta_2 \times V_2^T \\
&= \begin{pmatrix} 0,2670 & -0,2567 \\ 0,7486 & -0,3981 \\ 0,2670 & -0,2567 \\ 0,1182 & -0,0127 \\ 0,5198 & 0,8423 \\ 0,1182 & -0,0127 \end{pmatrix} \times \begin{pmatrix} 1,6950 & 0 \\ 0 & 1,1158 \end{pmatrix} \times \begin{pmatrix} 0,4366 & 0,3067 & 0,4412 & 0,4909 & 0,5288 \\ -0,4717 & 0,7549 & -0,3568 & -0,0346 & 0,2815 \end{pmatrix} \\
&= \begin{pmatrix} 0,3327 & -0,0774 & 0,3019 & 0,2321 & 0,1587 \\ 0,7630 & 0,0535 & 0,7178 & 0,6377 & 0,5453 \\ 0,3327 & -0,0774 & 0,3019 & 0,2321 & 0,1587 \\ 0,0941 & 0,0507 & 0,0934 & 0,0988 & 0,1019 \\ -0,0586 & 0,9867 & 0,0534 & 0,4000 & 0,7305 \\ 0,0808 & 0,0722 & 0,0833 & 0,0978 & 0,1099 \end{pmatrix} \qquad (25)
\end{aligned}
$$

Die Vektoren der beiden Anfragen $q_1 =$ *baking* and $q_2 =$ *baking bread* und ihre Projektionen auf den Themenraum $\vec{q}' = \vec{q} \times U_2$ sind folgendermaßen:

$$
\begin{aligned}
\vec{q_1} &= \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} && (26) \\
\vec{q_2} &= \begin{pmatrix} 1 & 0 & 1 & 0 & 0 & 0 \end{pmatrix} && (27) \\
\vec{q_1}' &= \begin{pmatrix} 0,2670 & -0,2567 \end{pmatrix} && (28) \\
\vec{q_2}' &= \begin{pmatrix} 0,5340 & -0,5134 \end{pmatrix} && (29)
\end{aligned}
$$

Jetzt erhalten wir die Ähnlichkeiten:

$$
\begin{aligned}
sim\,(q_1, d_1) &= \vec{q_1}' \times {V_2^T}_{*1} \approx 0,2377 && (30) \\
sim\,(q_1, d_2) &= \vec{q_1}' \times {V_2^T}_{*2} \approx -0,1119 && (31) \\
sim\,(q_1, d_3) &= \vec{q_1}' \times {V_2^T}_{*3} \approx 0,2094 && (32) \\
sim\,(q_1, d_4) &= \vec{q_1}' \times {V_2^T}_{*4} \approx 0,1400 && (33) \\
sim\,(q_1, d_5) &= \vec{q_1}' \times {V_2^T}_{*5} \approx 0,0689 && (34)
\end{aligned}
$$

$$
\begin{aligned}
sim\,(q_2, d_1) &= \vec{q_2}' \times {V_2^T}_{*1} \approx 0,4753 && (35) \\
sim\,(q_2, d_2) &= \vec{q_2}' \times {V_2^T}_{*2} \approx -0,2238 && (36) \\
sim\,(q_2, d_3) &= \vec{q_2}' \times {V_2^T}_{*3} \approx 0,4188 && (37) \\
sim\,(q_2, d_4) &= \vec{q_2}' \times {V_2^T}_{*4} \approx 0,2869 && (38) \\
sim\,(q_2, d_5) &= \vec{q_2}' \times {V_2^T}_{*5} \approx 0,1386 && (39)
\end{aligned}
$$

... und die beiden (identischen) Rankings:

$$
\begin{aligned}
Ranking_1 &= d_1, d_3, d_4, d_5, d_2 && (40) \\
Ranking_2 &= d_1, d_3, d_4, d_5, d_2 && (41)
\end{aligned}
$$