Übungen zu Multimedia-Datenbanken
Aufgabenblatt 2 - Musterlösungen

Übung: Dipl.-Inform. Tina Walber
Vorlesung: Dr.-Ing. Marcin Grzegorzek
Fachbereich Informatik, Universität Koblenz–Landau

Ausgabe:	03.05.2010
Abgabe:	16.05.2010 per Email an **walber@uni-koblenz.de** als PDF-Anhang
Format:	mmdb-blatt2-nachname1-nachname2.pdf

1 Information Retrieval - Boole'sches Modell (4 Punkte)

Gegeben seien Dokumente $d_1, \ldots, d_6$ und die Indexmenge[1] I={Eyjafjalla, Vulkan, Island, 2010, Satellitenbild, Aschewolke, Flugverbot, Athmosphäre}. Nehmt folgende Zuweisung von Indextermen zu Dokumenten an:

d_1	{Eyjafjalla,Vulkan,2010,Aschewolke}
d_2	{Eyjafjalla,2010}
d_3	{Island,Eyjafjalla,Athmosphäre,Satellitenbild}
d_4	{Vulkan,Flugverbot,Aschewolke}
d_5	{2010,Vulkan,Flugverbot}
d_6	{Island,Vulkan,Flugverbot}

1. Schreibt eine Query in konjunktiver Normalform die alle Dokumente über Vulkan oder Eyjafjalla aus dem Jahr 2010 sowie über Aschewolke oder Flugverbot zurück liefert.
2. Schreibt die Query in disjunktive Normalform um.
3. Gebt die Liste der zurückgelieferten Dokumente an.
4. Was ist der Nachteil des boole'schen Modells? Warum ist es eher Daten- als Information-Retrieval?

Musterlösung:

1. Query in KNF (1 Punkt): *(Eyjafjalla or Vulkan) and 2010 and (Aschewolke or Flugverbot)*
2. Query in DNF (1 Punkt): *(Eyjafjalla and 2010 and Aschewolke) or (Eyjafjalla and 2010 and Flugverbot) or (Vulkan and 2010 and Aschewolke) or (Vulkan and 2010 and Flugverbot)*

[1] Zur Info: Der Vulkan Eyjafjallajökull wird des öfteren in Nachrichtenmagazinen nur Eyjafjalla genannt. Der Einfachheit halber haben wir diese Bezeichnung übernommen.

3. (d_1, d_5) (1 Punkt)

4. Das boole'sche Modell ist ein **exaktes Modell**. Im Information Retrieval steht jedoch die Informationsuche im Vordergrund, wobei die **Information aber oft nur ungenau spezifiziert** werden kann. Dadurch muss ein IR Modell mit inexakten Daten umgehen können und gerade das Zurückliefern von Dokumenten anhand ihrer **Relevanz** unterstützen. **Ähnlichkeitssuche** wird durch exakte Anfrageformulierung **nicht wirklich unterstützt**. (1 Punkt)

---

2 Information Retrieval - Fuzzy Modell (6 Punkte)

Gegeben seien nochmal die Dokumente $d_1, \ldots, d_6$ von Aufgabe 1 mit der reduzierten Indexmenge I={Eyjafjalla, Vulkan, Island, Aschewolke, Flugverbot}. Nehmt folgende Zuweisung von Indextermen zu Dokumenten an:

d_1 {Eyjafjalla,Vulkan,Aschewolke}
d_2 {Eyjafjalla}
d_3 {Island,Eyjafjalla}
d_4 {Vulkan,Flugverbot,Aschewolke}
d_5 {Vulkan,Flugverbot}
d_6 {Island,Vulkan,Flugverbot}

1. Berechnet die Korrelationsmatrix für die Indexterme anhand des Ansatzes von Ogawa, Morita und Kobayashi. Gebt sinnvolle Zwischenschritte bei der Rechnung an.

2. Berechnet die Zugehörigkeit der Indexterme zu den Dokumenten, und stellt die Ergebnisse tabellarisch dar. Gebt sinnvolle Zwischenschritte bei der Rechnung an.

3. Berechnet die folgenden Queries anhand des Fuzzy Modells und gebt eine anhand der Relevanz sortierte Ergebnisliste an.

   a) *Vulkan and Eyjafjalla*

   b) *not Aschewolke*

Musterlösung:

1. (2 Punkte)

	Eyjafjalla	Vulkan	Island	Aschewolke	Flugverbot
Eyjafjalla	1	0.17	0.25	0.25	0
Vulkan	0.17	1	0.2	0.5	0.75
Island	0.25	0.2	1	0	0.25
Aschewolke	0.25	0.5	0	1	0.25
Flugverbot	0	0.75	0.25	0.25	1

Rechnung Die Korrelation zwischen den Termen t_i und t_j berechnet sich nach der Formel $c_{ij} = \frac{n_{ij}}{n_i+n_j-n_{ij}}$, mit n_{ij} Anzahl Dokumente mit Indextermen t_i und t_j, n_i Anzahl Auftreten t_i und analog für n_j. Es gilt $c(t,t) = 1$, ein Term t korreliert zu sich selbst mit 1, ausserdem $c(t_i,t_j) = c(t_j,t_i)$. Die entsprechenden Rechnungen werden daher nicht explizit aufgeführt.

$c(Eyjafjalla, Vulkan) = \frac{1}{3+4-1} = 0.17$
$c(Eyjafjalla, Island) = \frac{1}{3+2-1} = 0.25$
$c(Eyjafjalla, Aschewolke) = \frac{1}{3+2-1} = 0.25$
$c(Eyjafjalla, Flugverbot) = \frac{0}{3+3-0} = 0$
$c(Vulkan, Island) = \frac{1}{4+2-1} = 0.2$
$c(Vulkan, Aschewolke) = \frac{2}{4+2-2} = 0.5$
$c(Vulkan, Flugverbot) = \frac{3}{4+3-3} = 0.75$
$c(Island, Aschewolke) = \frac{0}{2+2-0} = 0$
$c(Island, Flugverbot) = \frac{1}{2+3-1} = 0.25$
$c(Aschewolke, Flugverbot) = \frac{1}{2+3-1} = 0.25$

2. (2 Punkte)

	d1	d2	d3	d4	d5	d6
Eyjafjalla	1	1	1	0.38	0.17	0.38
Vulkan	1	0.17	0.34	1	1	1
Island	0.4	0.25	1	0.4	0.4	1
Aschewolke	1	0.25	0.25	1	0.63	0.63
Flugverbot	0.81	0	0.25	1	1	1

Rechnung Die Zugehörigkeit des Termes t_i zum Dokument d_j berechnet sich nach der Formel $\mu_{t_i}(d_j) = 1 - \prod_{t_k \in d_j}(1 - c_{ik})$. Offensichtlich gilt: Wenn $t_i \in d_j$. dann $\mu_{t_i}(d_j) = 1$, daher wird die Rechnung in diesen Fällen nicht aufgeführt. Im Folgenden werden die Indexterme durch den ersten Bunchstaben abgekürzt.

$\mu_E(d_4) = 1 - (1 - 0.17)(1 - 0.25)(1 - 0) = 1 - 0.83 * 0.75 * 1 = 0.38$
$\mu_E(d_5) = 1 - (1 - 0.17)(1 - 0) = 1 - 0.83 * 1 = 0.17$
$\mu_E(d_6) = 1 - (1 - 0.25)(1 - 0.17)(1 - 0) = 1 - 0.75 * 0.83 * 1 = 0.38$
$\mu_V(d_2) = 1 - (1 - 0.17) = 1 - 0.83 = 0.17$
$\mu_V(d_3) = 1 - (1 - 0.2)(1 - 0.17) = 1 - 0.8 * 0.83 = 0.34$
$\mu_I(d_1) = 1 - (1 - 0.25)(1 - 0.2)(1 - 0) = 1 - 0.75 * 0.8 * 1 = 0.4$
$\mu_I(d_2) = 1 - (1 - 0.25) = 1 - 0.75 = 0.25$
$\mu_I(d_4) = 1 - (1 - 0.2)(1 - 0.25)(1 - 0) = 1 - 0.8 * 0.75 * 1 = 0.4$
$\mu_I(d_5) = 1 - (1 - 0.2)(1 - 0.25) = 1 - 0.8 * 0.75 = 0.4$
$\mu_A(d_2) = 1 - (1 - 0.25) = 1 - 0.75 = 0.25$
$\mu_A(d_3) = 1 - (1 - 0)(1 - 0.25) = 1 - 1 * 0.75 = 0.25$
$\mu_A(d_5) = 1 - (1 - 0.5)(1 - 0.25) = 1 - 0.5 * 0.75 = 0.63$
$\mu_A(d_6) = 1 - (1 - 0)(1 - 0.5)(1 - 0.25) = 1 - 1 * 0.5 * 0.75 = 0.63$
$\mu_F(d_1) = 1 - (1 - 0)(1 - 0.75)(1 - 0.25) = 1 - 1 * 0.25 * 0.75 = 0.81$
$\mu_F(d_2) = 1 - (1 - 0) = 1 - 1 = 0$
$\mu_F(d_3) = 1 - (1 - 0.25)(1 - 0) = 1 - 0.75 * 1 = 0.25$

3. (je 1 Punkt pro Query. (2P))

   a) $\mu_{V \wedge E}(d_j) = \min(\mu_V(d_j), \mu_E(d_j))$

   $\mu_{V \wedge E}(d_1) = \min(\mu_V(d_1), \mu_E(d1)) = \min(1, 1) = 1$
   $\mu_{V \wedge E}(d_2) = 0.17$
   $\mu_{V \wedge E}(d_3) = 0.34$
   $\mu_{V \wedge E}(d_4) = 0.38$
   $\mu_{V \wedge E}(d_5) = 0.17$
   $\mu_{V \wedge E}(d_6) = 0.38$

   Die sortierte Ergebnisliste: $\{d_1, d_4, d_6, d_3, d_2, d_5\}$

   b) $\mu_{\neg A}(d_j) = 1 - \mu_A(d_j)$

   $\mu_{\neg A}(d_1) = 1 - \mu_R(d_1) = 1 - 1 = 0$
   $\mu_{\neg A}(d_2) = 0.75$
   $\mu_{\neg A}(d_3) = 0.75$
   $\mu_{\neg A}(d_4) = 0.0$
   $\mu_{\neg A}(d_5) = 0.37$
   $\mu_{\neg A}(d_6) = 0.37$

   Die sortierte Ergebnisliste: $\{d_2, d_3, d_5, d_6, d_1, d_4\}$

---

3 Information Retrieval - Vektorraummodell (6 Punkte)

Gegeben seien wieder die Dokumente aus Aufgabe 2. Betrachtet die Zugehörigkeitswerte aus Aufgabe 2.2 als Termgewichte. Gegeben sei die Query *Vulkan and Eyjafjalla*. Desweiteren sei eine Query durch das Dokument d_6 spezifiziert (Ähnlichkeitssuche).

1. Berechnet die Ähnlichkeit der Dokumente zu den beiden Queries mit Hilfe der euklidischen Distanz. Gebt eine nach Relevanz sortierte Ergebnisliste an. Gebt sinnvolle Zwischenschritte bei der Rechnung an.
2. Berechnet die Ähnlichkeit der Dokumente zu den beiden Queries mit Hilfe des Kosinusmaßes. Gebt eine nach Relevenz sortierte Ergebnisliste an. Gebt sinnvolle Zwischenschritte bei der Rechnung an.
3. Was sind die Nachteile bzw. Probleme des Vektorraumodells.

Musterlösung:

1. (2 Punkte)

   Die erste Query lautet als Vektor $q_1 = (1, 1, 0, 0, 0)$, und für die zweite gilt $q_2 = d_6$
   Die euklidische Distanz zwischen einem Dokument und einer Query ist definiert als:

   $dissim(d, q) = \sqrt{\sum_i (d[i] - q[i])^2}, 0 \leq i \leq |I|$

Jeder Indexterm wird also als eigene Dimension aufgefasst. In unserem Beispiel sollen die Zugehörigkeitswerte als Termgewichte betrachtet werden, es gilt also $d_j[i] = \mu_{t_i}(d_j)$.

$$
\begin{aligned}
dissim(d_1, q_1) &= \sqrt{\sum_{i=1\ldots5} (d_1[i] - q_1[i])^2} \\
&= \sqrt{(1-1)^2 + (1-1)^2 + (0.4-0)^2 + (1-0)^2 + (0.81-0)^2} \\
&= \sqrt{0 + 0 + 0.16 + 1 + 0.66} \\
&= 1.35 \\
dissim(d_2, q_1) &= \sqrt{(1-1)^2 + (0.17-1)^2 + (0.25-0)^2 + (0.25-0)^2 + (0-0)^2} \\
&= 0.9 \\
dissim(d_3, q_1) &= \sqrt{(1-1)^2 + (0.34-1)^2 + (1-0)^2 + (0.25-0)^2 + (0.25-0)^2} \\
&= 1.25 \\
dissim(d_4, q_1) &= \sqrt{(0.38-1)^2 + (1-1)^2 + (0.4-0)^2 + (1-0)^2 + (1-0)^2} \\
&= \sqrt{0.38 + 0 + 0.16 + 1 + 1} \\
&= 1.6 \\
dissim(d_5, q_1) &= \sqrt{(0.17-1)^2 + (1-1)^2 + (0.4-0)^2 + (0.63-0)^2 + (1-0)^2} \\
&= \sqrt{0.69 + 0 + 0.16 + 0.4 + 1} \\
&= 1.5 \\
dissim(d_6, q_1) &= \sqrt{(0.38-1)^2 + (1-1)^2 + (1-0)^2 + (0.63-0)^2 + (1-0)^2} \\
&= 1.67
\end{aligned}
$$

Ergebnisliste: $\{d_2, d_3, d_1, d_5, d_4, d_6\}$.

$dissim(d_1, q_2) = 0.96$
$dissim(d_2, q_2) = 1.67$
$dissim(d_3, q_2) = 1.24$
$dissim(d_4, q_2) = 0.7$
$dissim(d_5, q_2) = 0.64$
$dissim(d_6, q_2) = 0.0$

Ergebnisliste: $\{d_6, d_5, d_4, d_1, d_3, d_2\}$.

2. (2 Punkte)

Das Kosinusmass ist definiert als $sim_{\cos}(d, q) = \frac{\langle d,q \rangle}{\|d\| * \|q\|}$. Mit $\langle d, q \rangle = \sum_i d[i] * q[i]$ und $|d| = \sqrt{\langle d, d \rangle}$.

Query 1:

$$
\begin{aligned}
sim_{\cos}(d_1, q_1) &= \frac{\langle d_1, q_1 \rangle}{\|d_1\| * \|q_1\|} \\
&= \frac{1*1+1*1+0,4*0+1*0+0,8125*0}{\sqrt{(1^2+1^2+0,4^2+1^2+0,8125^2)*(1^2+1^2+0^2+0^2+0^2)}} \\
&= \frac{2}{\sqrt{3.82015625*2}} \\
&= \frac{2}{\sqrt{7.6403125}} \\
&= 0.72 \\
sim_{\cos}(d_2, q_1) &= 0.77 \\
sim_{\cos}(d_3, q_1) &= 0.63 \\
sim_{\cos}(d_4, q_1) &= 0.54 \\
sim_{\cos}(d_5, q_1) &= 0.51 \\
sim_{\cos}(d_6, q_1) &= 0.52
\end{aligned}
$$

Ergebnisliste: $\{d2, d1, d3, d4, d6, d5\}$.

Query 2:

$sim_{\cos}(d_1, q_2) = 0.88$
$sim_{\cos}(d_2, q_2) = 0.47$
$sim_{\cos}(d_3, q_2) = 0.76$
$sim_{\cos}(d_4, q_2) = 0.93$
$sim_{\cos}(d_5, q_2) = 0.95$
$sim_{\cos}(d_6, q_2) = 1$

Ergebnisliste: $\{d6, d5, d4, d1, d3, d2\}$.

3. (2 Punkte)
   - Merkmale werden als orthogonal aufgefasst, was unrealisitisch ist.
   - Probleme bzgl. Effizienz möglich bei vielen Merkmalen.
   - Keine Junktoren, da Anfrage ein Vektor ist.

---

4 Relevance Feedback (6 Punkte)

1. Was bedeutet Relevance Feedback?
2. Basierend auf den Dokumenten und den Indextermen aus Aufgabe 2: Ein User hat Dokument d_6 als relevant und d_2 als irrelevant eingestuft. Berechnet den neuen Anfragevektor für die Anfrage q=*Eyjafjalla and Vulkan* mit $\alpha = 1$ und $\beta = 0.5$ mit Hilfe des Verfahrens von Rocchio. Gebt sinnvolle Zwischenschritte bei der Rechnung an.
3. Berechnet die modifizierte Anfrage mit Hilfe des euklidischen Distanzmaßes. Hat sich eine Änderung gegenüber Aufgabe 3.1 ergeben, und wenn ja, wie kann man sie deuten? Gebt sinnvolle Zwischenschritte bei der Rechnung an.

Musterlösung:

1. **Bewertung von Ergebnisdokumenten** und anschließende **Neuberechnung der Anfrage** unter Zuhilfenahme der Bewertung. (2 Punkte)
2. Die Bewertung fließt mit Hilfe folgender Formel in die Originalquery ein: $q_{neu} = q_{alt} + \frac{\alpha}{|D_r|} \sum_{d_r \in D_r} (d_r) - \frac{\beta}{|D_i|} \sum_{d_i \in D_i} (d_i)$.

$$q_{neu} = \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} + 1 * \begin{pmatrix} 0.38 \\ 1 \\ 1 \\ 0.63 \\ 1 \end{pmatrix} - 0.5 * \begin{pmatrix} 1 \\ 0.17 \\ 0.25 \\ 0.25 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} + \begin{pmatrix} 0.38 \\ 1 \\ 1 \\ 0.63 \\ 1 \end{pmatrix} - \begin{pmatrix} 0.5 \\ 0.085 \\ 0.125 \\ 0.125 \\ 0 \end{pmatrix} = \begin{pmatrix} 0.88 \\ 1.915 \\ 0.875 \\ 0.505 \\ 1 \end{pmatrix} = \begin{pmatrix} \frac{7}{8} \\ \frac{23}{12} \\ \frac{7}{8} \\ \frac{1}{2} \\ 1 \end{pmatrix}$$

(2 Punkte)

3. (1 Punkt für Liste, 1 Punkt für Erklärung)

$$dissim(d_1, q_{neu}) = 1.167$$
$$dissim(d_2, q_{neu}) = 2.126$$
$$dissim(d_3, q_{neu}) = 1.773$$
$$dissim(d_4, q_{neu}) = 1.247$$
$$dissim(d_5, q_{neu}) = 1.256$$
$$dissim(d_6, q_{neu}) = 1.057$$

Ergebnisliste: $\{d_6, d_1, d_4, d_5, d_3, d_2\}$.

Das als relevant eingestufte Dokument ist deutlich höher gerankt, und das als irrelevant eingestufte Dokument ans Ende der Liste gewandert.

---

5 Bewertung von Retrieval Modellen (8 Punkte)

1. Erläutert Precision, Recall und Fall-Out. Gebt auch die jeweilige Berechnungsvorschrift an.

2. Gegeben seinen Dokumente $d_1, \ldots, d_{20}$. Bezüglich einer Anfrage q seinen die Dokumente $\{d_2, d_5, d_9, d_{11}, d_{14}\}$ relevant. Zwei Systeme geben die Ergebnisliste $e_1 := \{d_2, d_4, d_5, d_9\}$ und $e_2 := \{d_2, d_3, d_5, d_6, d_8, d_9, d_{11}, d_{12}\}$. Berechnet Precision, Recall, Fall-Out.

3. Wie unterscheiden die beiden Systeme sich in ihrem Verhalten?

Musterlösung:

1. (Je 0.5 Punkte für Erklärung und 0.5 Punkte für Formel (3P).)

   Precision Beschreibt das Verhältnis der gefundenen und relevanten Dokumente zu allen gefundenen Dokumenten: $p = \frac{ca}{ca+fa}$.

   Recall Beschreibt das Verhältnis der gefundenen und relevanten Dokumente zu allen relevanten Dokumenten: $r = \frac{ca}{ca+fd}$.

   Fall-Out Beschreibt das Verhältnis der zurukgelieferten irrelevanten Dokumente zu allen irrelevanten Dokumenten: $f = \frac{fa}{fa+cd}$.

2. (0,5 Punkte pro richtigem Wert (3 Punkte))

	fa	ca	fd	cd	p	r	f
e_1	1	3	2	14	0.75	0.6	0.07
e_2	4	4	1	11	0.5	0.8	0.27

3. Das erste System liefert weniger Dokumente zurück, die dafür aber relevant sind. Daher hat es einen relativ hohen Precision-Wert, während Recall geringer ausfällt, und Fall-Out extrem klein ist. Das System liefert also wenig irrelevantes zurück.

Das zweite System liefert mehr zurück, und das Ergebnis enthält entsprechend mehr irrelevante Dokumente. Das schlägt sich in einer kleineren Precision und größerem Fall-Out nieder. Recall ist dafür sehr hoch. In diesem speziellen Fall scheint System 2 eher die "bessere" Antwort zu geben, bei größeren Ergebnislisten ist eine niedrige Precision und ein hoher Fall-Out jedoch unter Umständen problematisch, da es schwer wird, die relevanten Dokumente in dem Ergebnis zu erkennen. (2 Punkte)
