Übungen zu Multimedia-Datenbanken
Aufgabenblatt 2

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

Ausgabe:	03.05.2010
Abgabe:	09.05.2010 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?

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

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

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*

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.

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.

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?