Multimedia Retrieval im WS 2011/2012

2. Prinzipien des Information Retrieval

Prof. Dr.-Ing. Marcin Grzegorzek
Juniorprofessur für Mustererkennung
Institut für Bildinformatik im Department ETI
Fakultät IV der Universität Siegen

17. und 24. Oktober 2011

Inhalte und Termine

2.1 Einführung
2.2 IR-Modelle
2.3 RF
2.4 Bewertung
2.5 Profile

1. Einführung

1.1 Grundlegende Begriffe

1.2 Suche in einem MMDBS

1.3 MMDBMS-Anwendungen 11.10.2011

2. Prinzipien des Information Retrieval

2.1 Einführung

2.2 Information-Retrieval-Modelle

2.3 Relevance Feedback

2.4 Bewertung von Retrieval-Systemen 17.10.2011

2.5 Nutzerprofile

Inhalte und Termine

2.1 Einführung
2.2 IR-Modelle
2.3 RF
2.4 Bewertung
2.5 Profile

3. Prinzipien des Multimedia Retrieval

3.1 Besonderheiten der Verwaltung und des Retrievals

3.2 Ablauf des Multimedia-Information-Retrievals

3.3 Daten eines Multimedia-Retrieval-Systems

3.4 Feature

3.5 Eignung verschiedener Retrieval-Modelle

3.6 Multimedia-Ähnlichkeitsmodell

4. Feature-Transformationsverfahren

4.1 Diskrete Fourier-Transformation

4.2 Diskrete Wavelet-Transformation

4.3 Karhunen-Loeve-Transformation

4.4 Latent Semantic Indexing und Singulärwertzerlegung

Inhalte und Termine

2.1 Einführung
2.2 IR-Modelle
2.3 RF
2.4 Bewertung
2.5 Profile

5. Distanzfunktionen

5.1 Eigenschaften und Klassifikation

5.2 Distanzfunktionen auf Punkten

5.3 Distanzfunktionen auf Binärdaten

5.4 Distanzfunktionen auf Sequenzen

5.5 Distanzfunktionen auf allgemeinen Mengen

6. Ähnlichkeitsmaße

6.1 Einführung

6.2 Distanz versus Ähnlichkeit

6.3 Grenzen von Ähnlichkeitsmaßen

6.4 Konkrete Ähnlichkeitsmaße

6.5 Aggregation von Ähnlichkeitswerten

6.6 Umwandlung von Distanzen in Ähnlichkeitswerte und Normierung

6.7 Partielle Ähnlichkeit

Inhalte und Termine

2.1 Einführung
2.2 IR-Modelle
2.3 RF
2.4 Bewertung
2.5 Profile

7. Effiziente Algorithmen und Datenstrukturen

7.1 Hochdimensionale Indexstrukturen

7.2 Algorithmen zur Aggregation von Ähnlichkeitswerten

8. Anfragebehandlung

8.1 Einführung

8.2 Konzepte der Anfragebehandlung

8.3 Datenbankmodell

8.4 Sprachen

9. Zusammenfassung

Überblick

2.1 Einführung
2.2 IR-Modelle
2.3 RF
2.4 Bewertung
2.5 Profile

2.1 Einführung

2.2 Information-Retrieval-Modelle

2.3 Relevance Feedback

2.4 Bewertung von Retrieval-Systemen

2.5 Nutzerprofile

Überblick

2.1 Einführung
2.2 IR-Modelle
2.3 RF
2.4 Bewertung
2.5 Profile

2.1 Einführung

2.2 Information-Retrieval-Modelle

2.3 Relevance Feedback

2.4 Bewertung von Retrieval-Systemen

2.5 Nutzerprofile

DB-, IR-, und MMDB-Systeme

2.1 Einführung
2.2 IR-Modelle
2.3 RF
2.4 Bewertung
2.5 Profile

- IR- und DB-Systeme verwalten Daten, unterscheiden sich jedoch erheblich im Zugriff auf die Daten.

- Datenbankanfrage ist scharf:

  select ISBN
  from Buch
  where Titel = "Multimedia-Datenbanken"

- IR-Anfrage ist in der Regel unscharf formuliert:
  Finde alle Text-Dokumente, die sich mit dem Thema "Multimedia-Datenbanken" beschäftigen.

- MMDB-Systeme kombinieren Konzepte von DB- und IR-Systemen.

Informationsbedarf in einem IR-System

2.1 Einführung
2.2 IR-Modelle
2.3 RF
2.4 Bewertung
2.5 Profile

Der Informationsbedarf in einem IR-System kann unterschiedlich verstanden werden:

- als Dokument:
  Liefere alle Text-Dokumente, die ähnlich zum Text-Dokument #0821 sind

- als Anfrage:
  Datenbank `and` (Bild `or` Video)

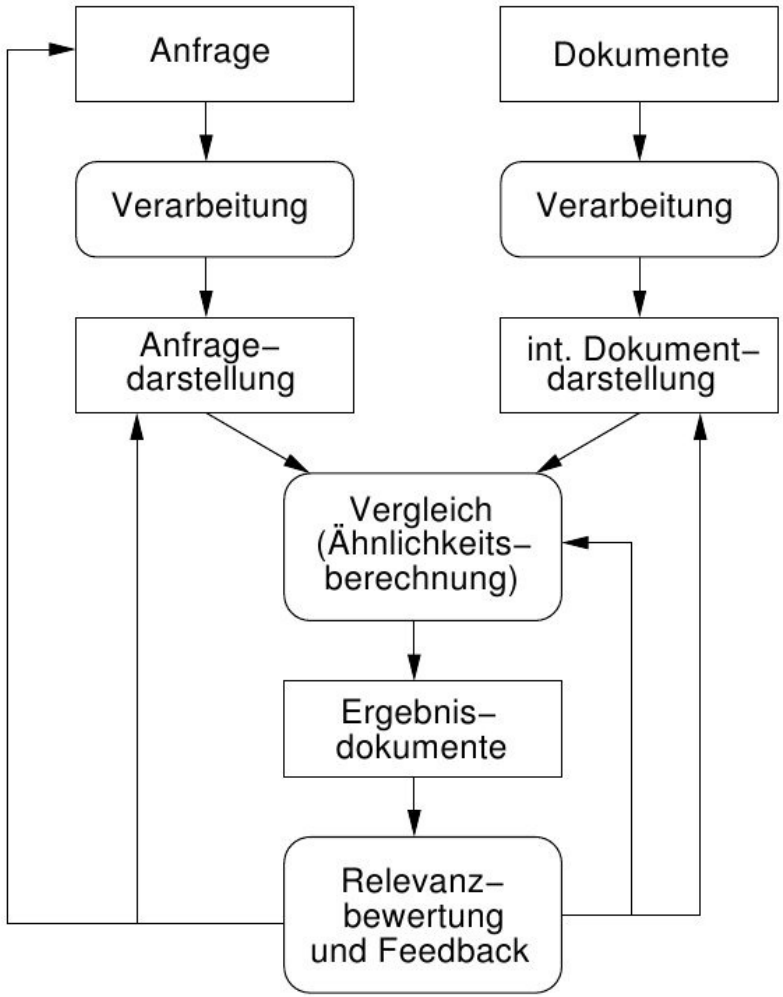
Daten Retrieval versus Information Retrieval

2.1 Einführung
2.2 IR-Modelle
2.3 RF
2.4 Bewertung
2.5 Profile

Merkmal	**Daten Retrieval**	**Inf. Retrieval**
Information	explizit	implizit
Ergebnisse	exakt	unscharf
Anfrage	einmalig	iterativ verfeinernd
Fehlertoleranz	keine	vorhanden
Ergebniskollektion	Menge	Liste

Schritte des IR-Prozesses

2.1 Einführung
2.2 IR-Modelle
2.3 RF
2.4 Bewertung
2.5 Profile



Überblick

2.1 Einführung
2.2 IR-Modelle
2.3 RF
2.4 Bewertung
2.5 Profile

2.1 Einführung

2.2 Information-Retrieval-Modelle

2.3 Relevance Feedback

2.4 Bewertung von Retrieval-Systemen

2.5 Nutzerprofile

Klassifikation der IR-Modelle

2.1 Einführung
2.2 IR-Modelle
2.3 RF
2.4 Bewertung
2.5 Profile

Boolesches Modell

- Dokumente werden als Indexterme repräsentiert.
- Die Suche erfolgt über einfache Mengenoperationen.
- Die Anfragen lassen sich über boolesche Junktoren verknüpfen.

Fuzzy-Modell

- Eine Erweiterung des booleschen Modells auf unscharfe Mengen.

Vektorraummodell

- Jedes Dokument wird als ein Vektor aufgefasst.
- Eine Anfrage wird auch als Vektor in einem Vektorraum behandelt.
- Die Suche basiert auf Bestimmung von Vektorähnlichkeiten.

Boolesches Modell - Allgemeines

2.1 Einführung
2.2 IR-Modelle
2.3 RF
2.4 Bewertung
2.5 Profile

Binäres Termgewicht

- Das Gewicht eines Terms bezogen auf ein Text-Dokument ist binär ("1" - das Dokument beinhaltet den Term, "0" - das Dokument beinhaltet den Term nicht).

Boolesche Junktoren

- In der Anfrage werden Terme durch boolesche Junktoren (and, or, not) kombiniert.

Vergleichsfunktion

- Innerhalb der Vergleichsfunktion werden die durch die Anfrage spezifizierten Anfrageterme in den jeweiligen Dokumenten auf Enthaltensein getestet.

Boolesches Modell - Beispiel

2.1 Einführung
2.2 IR-Modelle
2.3 RF
2.4 Bewertung
2.5 Profile

Terme:

Indexvokabular = {Korsika, Sardinien, Strand, Ferienwohnung, Gebirge}

Dokumente:

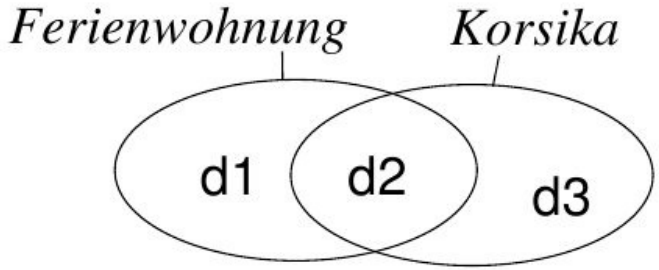
Dokument 1 :	{Sardinien, Strand, Ferienwohnung}
Dokument 2 :	{Korsika, Strand, Ferienwohnung}
Dokument 3 :	{Korsika, Gebirge}

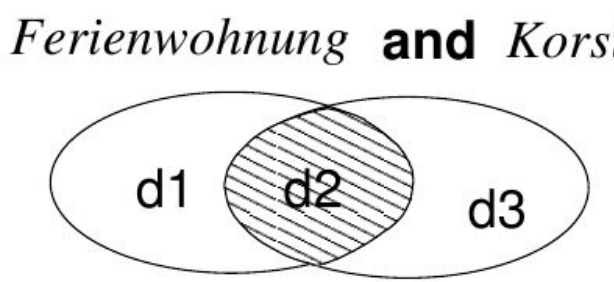
Ergebnisse:

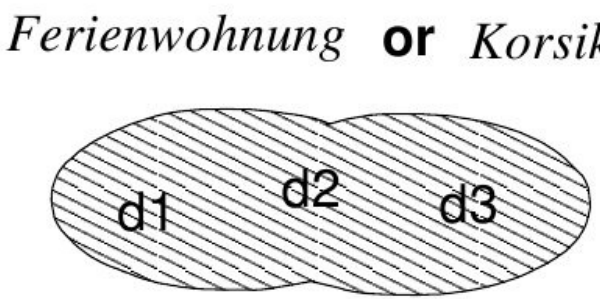
Korsika	liefert {d2,d3}
Ferienwohnung	liefert {d1,d2}
Ferienwohnung and Korsika	liefert {d2}
Ferienwohnung or Korsika	liefert {d1,d2,d3}
Ferienwohnung and not Korsika	liefert {d1}

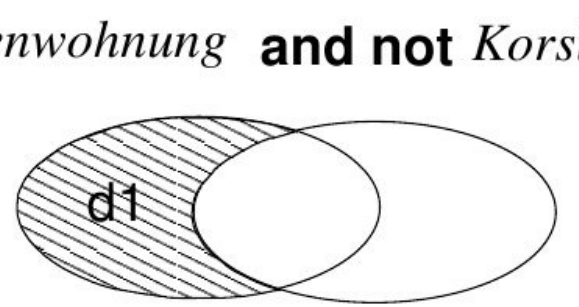
Boolesches Modell - Beispiel

2.1 Einführung
2.2 IR-Modelle
2.3 RF
2.4 Bewertung
2.5 Profile









Boolesches Modell - Weiteres

2.1 Einführung
2.2 IR-Modelle
2.3 RF
2.4 Bewertung
2.5 Profile

“but”-Junktor

- Die Anfrage *Ferienwohnung but Korsika* liefert alle die Dokumente, die den Term “Ferienwohnung” aber nicht den Term “Korsika” enthalten.

“of”-Konstrukt

- Die Anfrage *2 of (Korsika, Strand, Ferienwohnung)* sucht nach allen Dokumenten, die mindestens zwei der drei vorgegebenen Terme enthalten.

Fuzzy-Modell, Allgemeines

2.1 Einführung
2.2 IR-Modelle
2.3 RF
2.4 Bewertung
2.5 Profile

- Das Fuzzy-Modell ist eine Erweiterung des booleschen Modells.
- Das Problem der zu scharfen Enthaltenseinsbedingung von Termen in Dokumenten wird behoben.
- Die Grundidee liegt in der Verwendung einer graduellen Zugehörigkeit von Dokumenten zu Termen.
- Es wird auf das Konzept einer Fuzzy-Menge zurückgegriffen.

Fuzzy-Modell, Definitionen

2.1 Einführung
2.2 IR-Modelle
2.3 RF
2.4 Bewertung
2.5 Profile

Fuzzy-Menge
Eine Fuzzy-Menge $A = \{\langle u, \mu_A(u)\rangle\}$ über ein Universum U ist durch eine Zugehörigkeitsfunktion $\mu_A : U \to [0, 1]$ charakterisiert, welche jedem Element u des Universums U einen Wert $\mu_A(u)$ aus dem Intervall $[0, 1]$ zuordnet.

Term als Fuzzy-Menge
In unserem Retrieval-Szenario entspricht die Menge aller gespeicherten Dokumente dem Universum und ein Term einer Fuzzy-Menge.

Fuzzy-Wert
Fuzzy-Wert $\mu_t(d_1)$ des Dokuments d_1 bezüglich des Terms t drückt aus, wie stark der Term das Dokument charakterisiert.

Fuzzy-Modell, Definitionen

2.1 Einführung
2.2 IR-Modelle
2.3 RF
2.4 Bewertung
2.5 Profile

Mengendurchschnitt
Der Mengendurchschnitt $A \cap B$ (Konjunktion) wird durch die Min-Funktion realisiert $\mu_{A\cap B}(u) = \min(\mu_A(u), \mu_B(u))$

Mengenvereinigung
Die Mengenvereinigung $A \cup B$ (Disjunktion) wird durch die Max-Funktion realisiert $\mu_{A\cup B}(u) = \max(\mu_A(u), \mu_B(u))$

Komplementbildung
Die Komplementbildung $\overline{A}$ (Negation) bezüglich des Universums entspricht einer Subtraktion von 1
$\mu_{\overline{A}}(u) = 1 - \mu_A(u)$

Fuzzy-Modell, Beispiel

2.1 Einführung
2.2 IR-Modelle
2.3 RF
2.4 Bewertung
2.5 Profile

Anfrage	μ	d_1	d_2	d_3
	μ_{Korsika}	0,1	0,6	1
	μ_{Strand}	0,3	0,2	0,8
1	$\mu_{\text{Korsika} \cap \text{Strand}}$			
2	$\mu_{\text{Korsika} \cup \text{Strand}}$			
3	$\mu_{\overline{\text{Korsika}}}$			

Fuzzy-Modell, Beispiel

2.1 Einführung
2.2 IR-Modelle
2.3 RF
2.4 Bewertung
2.5 Profile

Anfrage	μ	d_1	d_2	d_3
	μ_{Korsika}	0,1	0,6	1
	μ_{Strand}	0,3	0,2	0,8
1	$\mu_{\text{Korsika} \cap \text{Strand}}$	0,1	0,2	0,8
2	$\mu_{\text{Korsika} \cup \text{Strand}}$	0,3	0,6	1
3	$\mu_{\overline{\text{Korsika}}}$	0,9	0,4	0

Fuzzy-Modell, Term-zu-Term-Korrelationsmatrix

2.1 Einführung
2.2 IR-Modelle
2.3 RF
2.4 Bewertung
2.5 Profile

- $c_{i,j}$ in Zeile t_i und Spalte t_j drückt aus, wie stark die Terme t_i und t_j in den Dokumenten der Datenbank korrelieren, also in den Dokumenten gemeinsam auftreten

$$c_{i,j} = \frac{n_{i,j}}{n_i + n_j - n_{i,j}}$$

- Zugehörigkeitswert eines Dokuments d_j zu einem Term t_j

$$\mu_{t_i}(d_j) = 1 - \prod_{t_k \in d_j} (1 - c_{i,k})$$

Vektorraummodell - Allgemeines

2.1 Einführung
2.2 IR-Modelle
2.3 RF
2.4 Bewertung
2.5 Profile

- Die Dokumente werden wie Vektoren eines Vektorraums aufgefasst.
- Vektoren aus Termgewichten oder Merkmalswerten
- Anfrage als Vektor
- Ähnlichkeits- und Distanzmaße

Vektorraummodell - Beispiel

2.1 Einführung
2.2 IR-Modelle
2.3 RF
2.4 Bewertung
2.5 Profile

Dokumente:

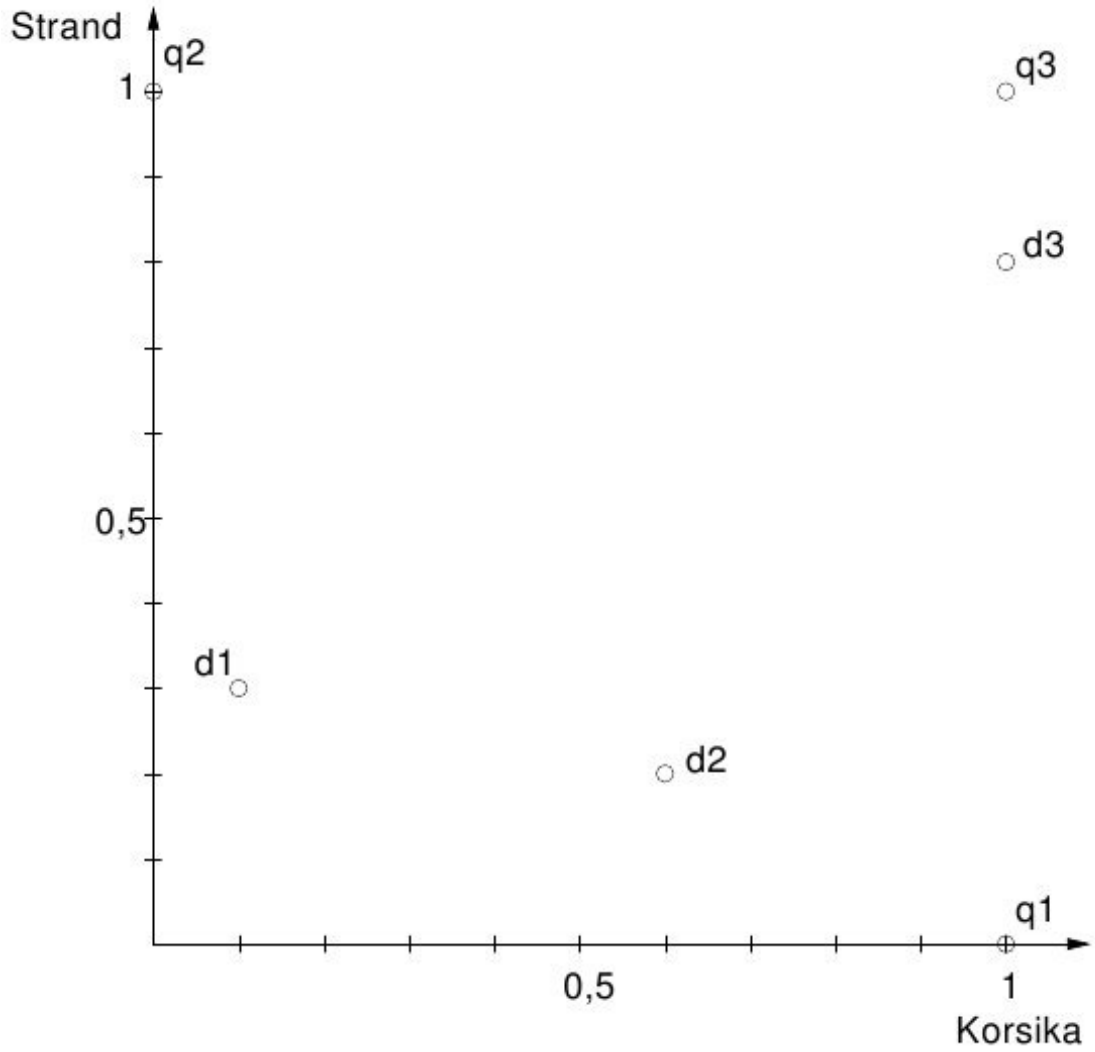
Dimension	d_1	d_2	d_3
Korsika	0,1	0,6	1
Strand	0,3	0,2	0.8

Anfragen:

Dimension	q_1	q_2	q_3
Korsika	1	0	1
Strand	0	1	1

Vektorraummodell - Beispiel

2.1 Einführung
2.2 IR-Modelle
2.3 RF
2.4 Bewertung
2.5 Profile



Vektorraummodell - Beispiel

2.1 Einführung
2.2 IR-Modelle
2.3 RF
2.4 Bewertung
2.5 Profile

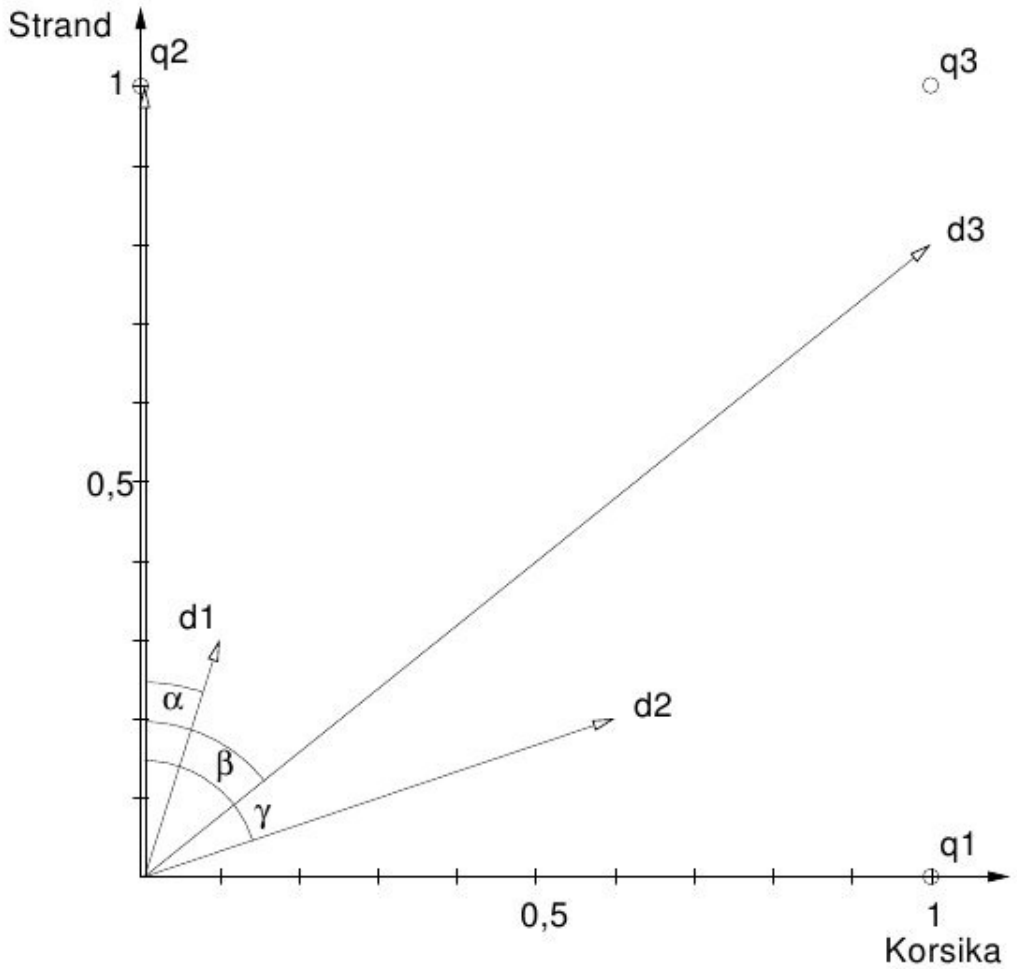
Ähnlichkeitswerte nach Kosinusmaß

$\mathrm{sim}_{\cos}$	d_1	d_2	d_3
q_1	0,3162	0,9487	0,7809
q_2	0,9487	0,3162	0,6247
q_3	0,8944	0,8944	0,9939

Vektorraummodell - Beispiel

2.1 Einführung
2.2 IR-Modelle
2.3 RF
2.4 Bewertung
2.5 Profile

$$\cos\alpha = 0,9487$$
$$\cos\beta = 0,6247$$
$$\cos\gamma = 0,3162$$



Vektorraummodell - Beispiel

2.1 Einführung
2.2 IR-Modelle
2.3 RF
2.4 Bewertung
2.5 Profile

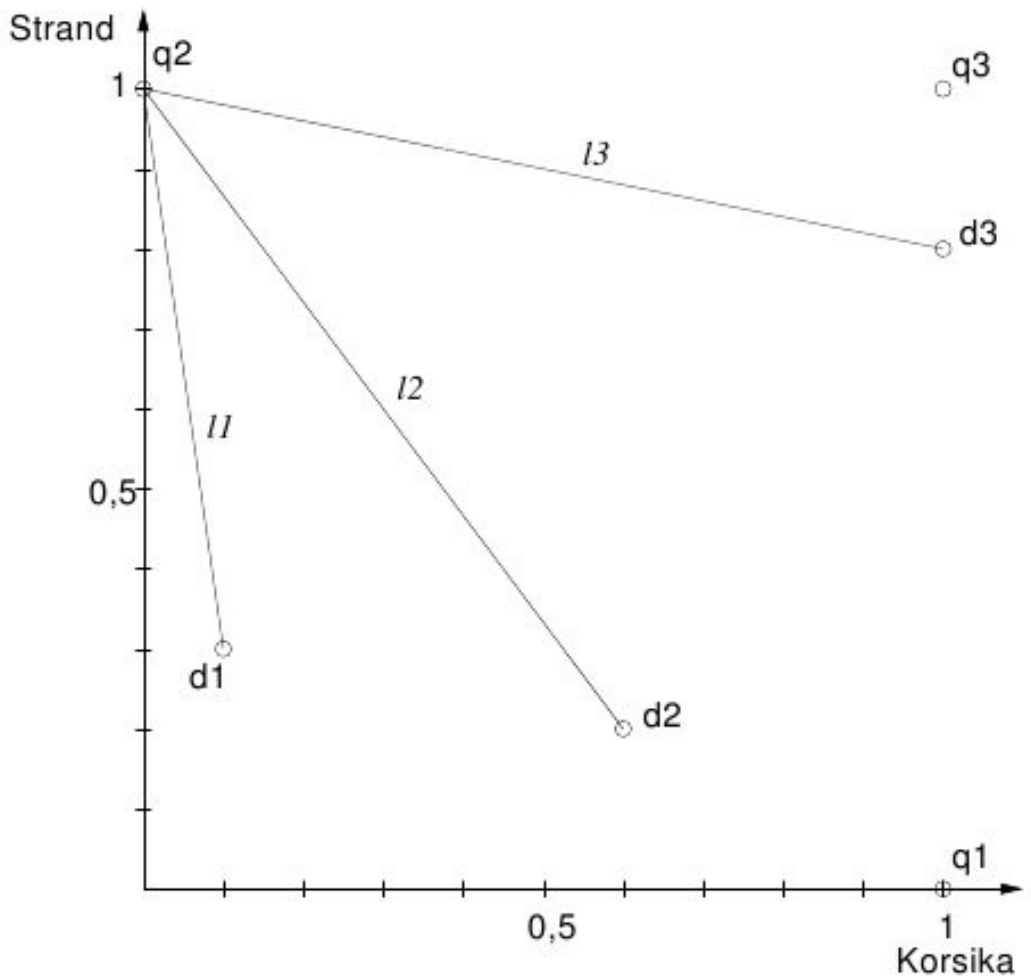
Unähnlichkeitswerte anhand Euklidischer Distanz

dissim_{L_2}	d_1	d_2	d_3
q_1	0,9487	0,4472	0,8
q_2	0,7071	1	1,0198
q_3	1,1402	0,8944	0,2

Vektorraummodell - Beispiel

2.1 Einführung
2.2 IR-Modelle
2.3 RF
2.4 Bewertung
2.5 Profile

$l_1 = 0,7071$
$l_2 = 1$
$l_3 = 1,0198$



Vektorraummodell - Beispiel

2.1 Einführung
2.2 IR-Modelle
2.3 RF
2.4 Bewertung
2.5 Profile

- Kosinusmaß und Euklidische Distanz erzeugen unterschiedliche Ergebnisse
  - Kosinusmaß erzeugt $\langle d_1, d_3, d_2 \rangle$
  - Euklidische Distanz erzeugt $\langle d_1, d_2, d_3 \rangle$

- Wahl der geeigneten Ähnlichkeitsfunktion abhängig von
  - subjektivem Ähnlichkeitsempfinden
  - Anwendungsszenario

Vektorraummodell - Zusammenfassung

2.1 Einführung
2.2 IR-Modelle
2.3 RF
2.4 Bewertung
2.5 Profile

- Vektorraummodell ist sehr verbreitet.
- VR-Modell setzt feste Anzahl von nummerischen Merkmalswerten pro Dokument voraus.
- Probleme:
  - Merkmale als orthogonale Dimensionen aufgefasst (unrealistisch)
  - Problem bei hoher Anzahl von Merkmalswerten bzgl. Effektivität und Effizienz
  - Anfrage ist Vektor, also keine Junktoren

Überblick

2.1 Einführung
2.2 IR-Modelle
2.3 RF
2.4 Bewertung
2.5 Profile

2.1 Einführung

2.2 Information-Retrieval-Modelle

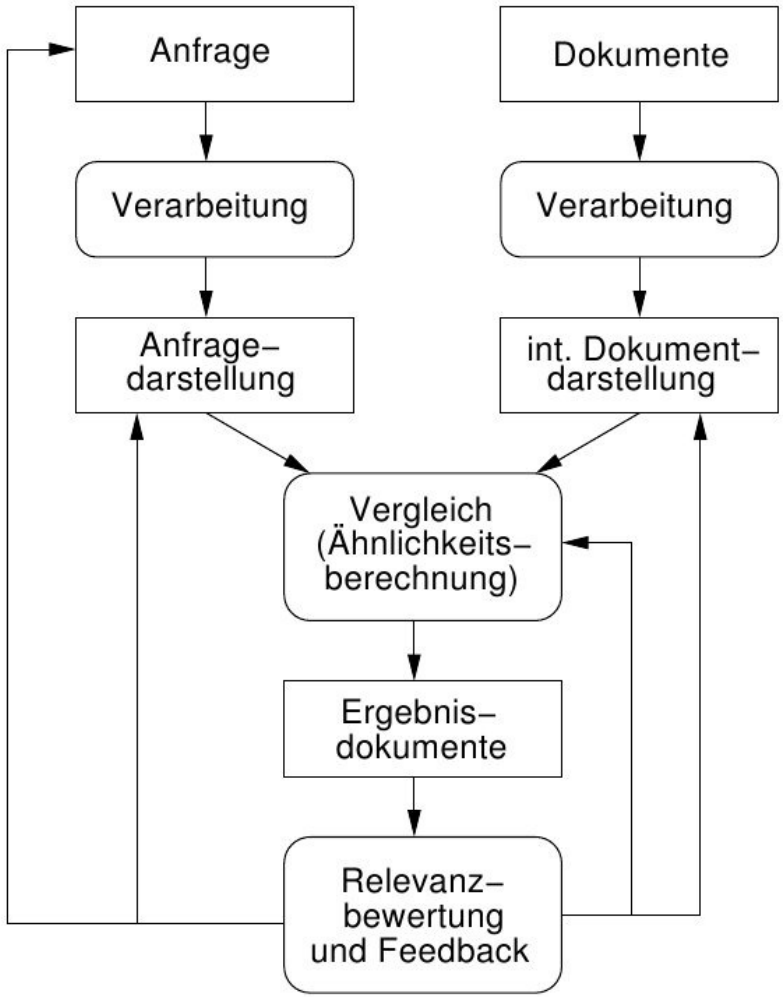
2.3 Relevance Feedback

2.4 Bewertung von Retrieval-Systemen

2.5 Nutzerprofile

Vereinfachter IR-Prozess

2.1 Einführung
2.2 IR-Modelle
2.3 RF
2.4 Bewertung
2.5 Profile



RF - Beispiel

2.1 Einführung
2.2 IR-Modelle
2.3 RF
2.4 Bewertung
2.5 Profile

Anfrage	**Ergebnisdokumente**			
	1	**2**	**3**	$\cdots$
q	d_0	d_1	d_2	$\cdots$
q_0	d_4	d_1 (+)	d_5 (−)	$\cdots$
q_1	d_1 (+)	d_3 (+)	d_4 (−)	$\cdots$
q_2	d_3	d_1	d_0	$\cdots$

q korrekte aber unbekannte Anfrage

q_0 initiale Anfrage

q_1 Anfrage nach 1. Iteration

q_2 Anfrage nach 2. Iteration

RF - Anfragemodifikation im Vektorraummodell

2.1 Einführung
2.2 IR-Modelle
2.3 RF
2.4 Bewertung
2.5 Profile

Verschiebung des Anfragevektors

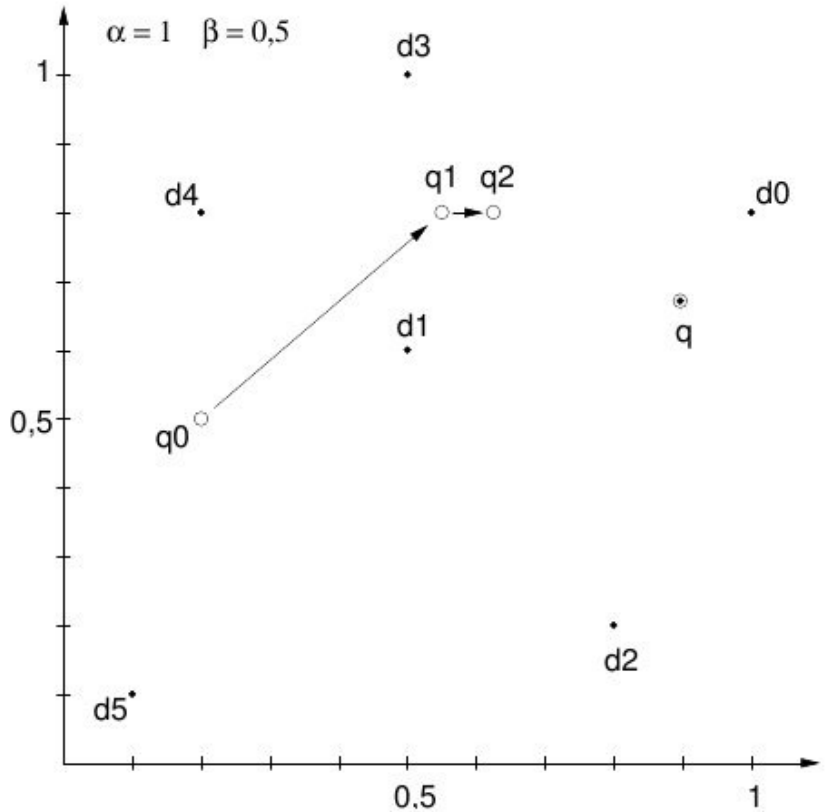
- in Richtung der als relevant bewerteten Dokumente
- weg von als irrelevant bewerteten Dokumenten
- Anfrage: q_{alt}, relevant (irrelevant) bewertete Dokumente: D_r (D_i)
- α und β gewichten Einfluss relevanter und irrelevanter Dokumente

$$q_{\text{neu}} = q_{\text{alt}} + \frac{\alpha}{|D_r|} \sum_{d_r \in D_r} d_r - \frac{\beta}{|D_i|} \sum_{d_i \in D_i} d_i$$

RF - Beispiel zur Anfragemodifikation

2.1 Einführung
2.2 IR-Modelle
2.3 RF
2.4 Bewertung
2.5 Profile

An-frage	Ergebnisdokumente			
	1	2	3	...
q	d_0	d_1	d_2	...
q_0	d_4	d_1 (+)	d_5 (−)	...
q_1	d_1 (+)	d_3 (+)	d_4 (−)	...
q_2	d_3	d_1	d_0	...



Überblick

2.1 Einführung
2.2 IR-Modelle
2.3 RF
2.4 Bewertung
2.5 Profile

2.1 Einführung

2.2 Information-Retrieval-Modelle

2.3 Relevance Feedback

2.4 Bewertung von Retrieval-Systemen

2.5 Nutzerprofile

Bewertung - Allgemeines

2.1 Einführung
2.2 IR-Modelle
2.3 RF
2.4 Bewertung
2.5 Profile

- Bewertung (Qualitätsvergleich) verschiedener Retrieval-Systeme
- Quantitative Maße vonnöten

Bewertung - Precision, Recall und Fallout

2.1 Einführung
2.2 IR-Modelle
2.3 RF
2.4 Bewertung
2.5 Profile

- Zwei verschiedene Fehlentscheidungen
  - *false alarms* (f_a) bezeichnet diejenigen Dokumente, die vom Retrieval-System irrtümlicherweise als relevant zurückgeliefert wurden (auch: *false positives*)
  - *false dismissals* (f_d) sind Dokumente, die fälschlicherweise vom Retrieval-System als irrelevant eingestuft wurden (auch: *false negatives*)

- Zwei korrekte Entscheidungen
  - *correct alarms* (c_a)
  - *correct dismissals* (c_d)

- f_a, f_d, c_a, c_d stehen für entsprechende Dokumentanzahlen bzgl. einer Anfrage.

Bewertung - Precision, Recall and Fallout

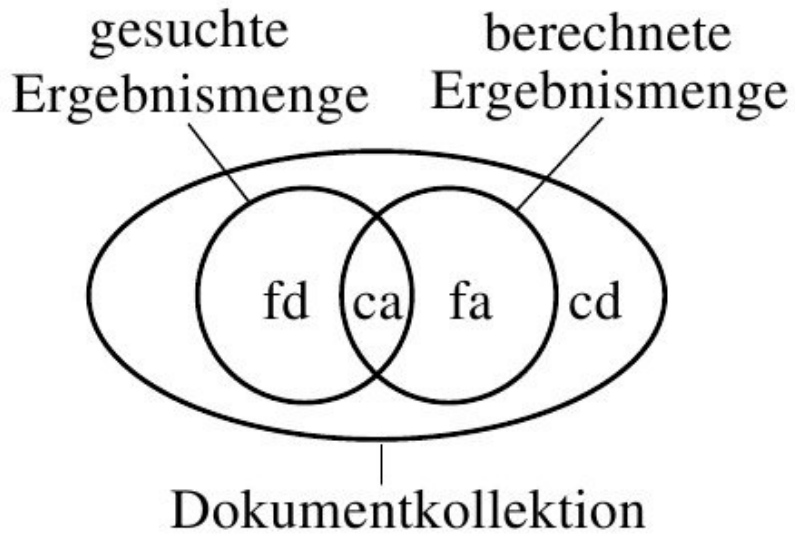
2.1 Einführung
2.2 IR-Modelle
2.3 RF
2.4 Bewertung
2.5 Profile

Nutzer-bewertung	Systembewertung	
	relevant	irrelevant
relevant	ca	fd
irrelevant	fa	cd

Bewertung - Precision, Recall and Fallout

2.1 Einführung
2.2 IR-Modelle
2.3 RF
2.4 Bewertung
2.5 Profile

$$
\begin{aligned}
|\text{gesuchte Ergebnismenge}| &= \text{fd} + \text{ca} \\
|\text{berechnete Ergebnismenge}| &= \text{ca} + \text{fa} \\
|\text{Dokumentkollektion}| &= \text{fd} + \text{ca} + \text{fa} + \text{cd}
\end{aligned}
$$



Bewertung - Precision

2.1 Einführung
2.2 IR-Modelle
2.3 RF
2.4 Bewertung
2.5 Profile

Precision P

Wie viele (als Verhältnis) Ergebnisdokumente sind tatsächlich relevant?

$$P = \frac{c_a}{c_a + f_a} \qquad P \in [0, 1]$$

Bewertung - Recall

2.1 Einführung
2.2 IR-Modelle
2.3 RF
2.4 Bewertung
2.5 Profile

Recall R

Wie viele (als Verhältnis) tatsächlich relevante Dokumente erscheinen im Ergebnis?

$$R = \frac{c_a}{c_a + f_d} \qquad R \in [0, 1]$$

Bewertung - Fallout

2.1 Einführung
2.2 IR-Modelle
2.3 RF
2.4 Bewertung
2.5 Profile

Fallout F

Verhältnis falsch gefundener zur Gesamtzahl irrelevanter Dokumente

$$F = \frac{f_a}{f_a + c_d} \qquad F \in [0, 1]$$

Bewertung - Precision, Recall und Fallout

2.1 Einführung
2.2 IR-Modelle
2.3 RF
2.4 Bewertung
2.5 Profile

- Precision, Recall und Fallout sind definiert bzgl. einer Anfrage
- Es ist besser, mehrere Anfragen zu betrachten und entsprechende Durchschnittswerte zu berechnen.

Bewertung - Precision, Recall and Fallout

2.1 Einführung
2.2 IR-Modelle
2.3 RF
2.4 Bewertung
2.5 Profile

20 Dokumente, 2 Anfragen, jeweils 10 Ergebnisdokumente

Anfrage	fa	ca	fd	cd	Precision	Recall	Fallout
q_1	8	2	6	4	20%	25%	66%
q_2	2	8	2	8	80%	80%	20%
Durchschnitt	–	–	–	–	50%	52,5%	43%

Überblick

2.1 Einführung
2.2 IR-Modelle
2.3 RF
2.4 Bewertung
2.5 Profile

2.1 Einführung

2.2 Information-Retrieval-Modelle

2.3 Relevance Feedback

2.4 Bewertung von Retrieval-Systemen

2.5 Nutzerprofile

Nutzerprofile - Allgemeines

2.1 Einführung
2.2 IR-Modelle
2.3 RF
2.4 Bewertung
2.5 Profile

- Bis jetzt keine Unterscheidung von Anwender und Anwendergruppen
- Verhalten bzw. Suchbedarf verschiedener Nutzer differiert oft
- Idee: Subjektivität wird als Nutzerprofil modelliert und bei Suche berücksichtigt

Nutzerprofile - Retrieval mit Profilen

2.1 Einführung
2.2 IR-Modelle
2.3 RF
2.4 Bewertung
2.5 Profile

Nachfiltern

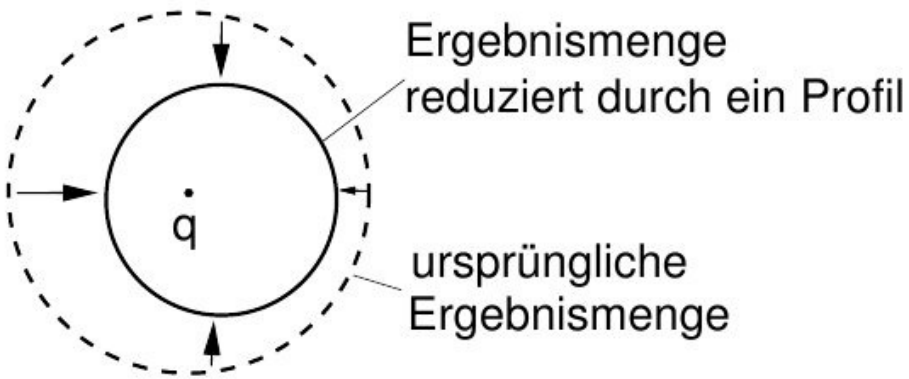
- Filterung auf Anfrageergebnis
- hoher Berechnungsaufwand durch u. U. großem Zwischenergebnis
- reduzieren von nur false alarms

Vorfiltern

- Nutzerprofil beeinflusst Retrieval-Prozess direkt
- Reduzierung von false alarms und false dismissals

Nutzerprofile - Nachfiltern

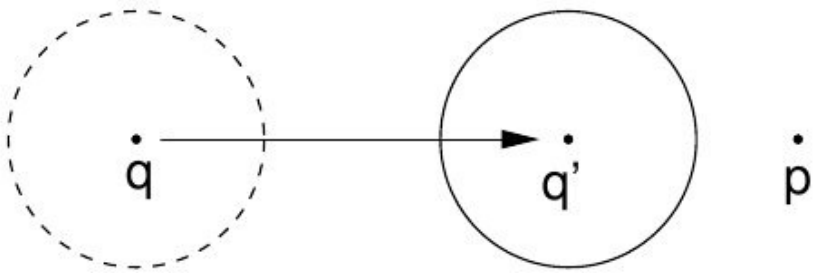
2.1 Einführung
2.2 IR-Modelle
2.3 RF
2.4 Bewertung
2.5 Profile



Nutzerprofile - Vorfiltern

2.1 Einführung
2.2 IR-Modelle
2.3 RF
2.4 Bewertung
2.5 Profile

- Annahme: Anfrage als Profil
- einfache Realisierung: Verschiebung Anfragepunkt q in Richtung Profilanfragepunkt p



- Problem: relevante Dokumente bzgl. q können irrelevant bzgl. q' werden
- gewünscht: Reduzierung false dismissals statt Reduzierung false alarms