

basiert auf Mengentheorie und Boole'scher Algebra  
sehr einfaches Modell mit klarer Semantik

Dokumente als Mengen von Indextermen → Termgewichte  
sind binär: im Dokument enthalten oder nicht enthalten

Test auf Enthaltensein als Vergleichsfunktion

Verknüpfung von Enthaltenseinsbedingungen mittels  
Boole'scher Junktoren

- ◆ *and* (Mengendurchschnitt)
- ◆ *or* (Mengenvereinigung)
- ◆ *not* (Mengendifferenz)

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

Dokument d1:{Sardinien, Strand, Ferienwohnung}

Dokument d2:{Korsika, Strand, Ferienwohnung}

Dokument d3:{Korsika, Gebirge}

### Anfrage

Korsika

liefert {d2, d3}

Ferienwohnung

liefert {d1, d2}

Ferienwohnung *and* Korsika

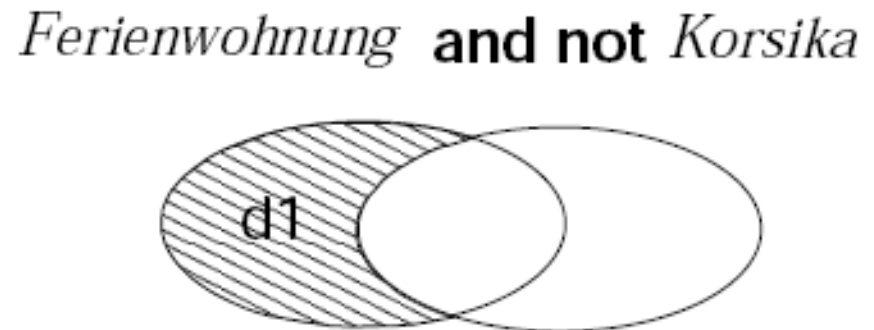
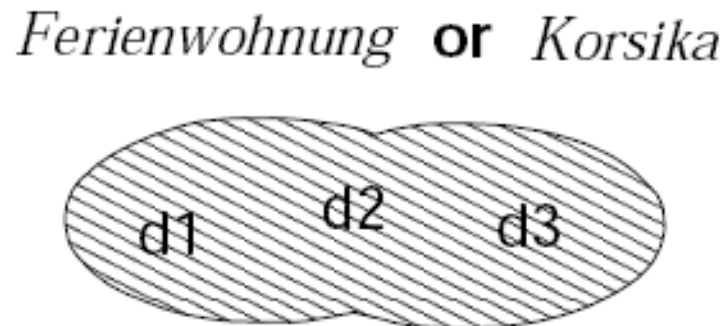
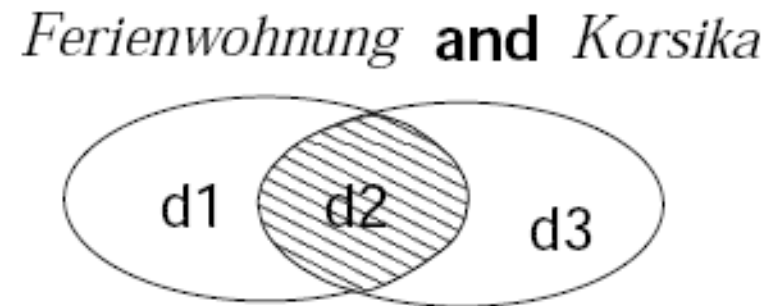
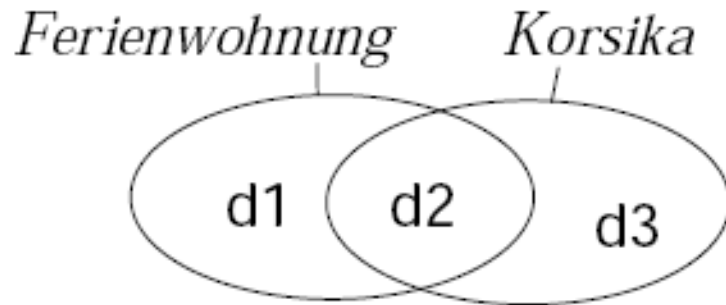
liefert {d2}

Ferienwohnung *or* Korsika

liefert {d1, d2, d3}

Ferienwohnung *and not* Korsika

liefert {d1}



## Anfragenormalisierung in DNF bzw. KNF

### Anfrage:

Ferienwohnung and ((Sardinien and Strand) or Korsika)

### *in disjunktiver Normalform (DNF):*

(Ferienwohnung and Sardinien and Strand) or  
(Ferienwohnung and Korsika)

### *in konjunktiver Normalform (KNF):*

Ferienwohnung and (Sardinien or Korsika) and (Strand or  
Korsika)

## Anfrageauswertung

jeder Term liefert Menge von Dokumenten, die diesen Term enthalten

komplexe Anfrage: Kombination der Dokumentenmengen durch entsprechende Mengenoperationen

kleine Zwischenergebnisse durch DNF (zuerst Durchschnitt, dann Vereinigung)

## Nachteile des Boole'schen Modells

exaktes Modell aufgrund binärer Gewichte

- ◆ eher Daten-Retrieval
- ◆ keine Ähnlichkeitssuche durch zu scharfe Suche

Größe des Ergebnisses

- ◆ alle Dokumente sind bzgl. Anfrage gleichrangig → Präsentation der gesamten Ergebnismenge
- ◆ Ergebnismenge in Abhängigkeit von Anfrage oft zu groß oder zu klein bzw. leer

## Nachteile des Boole'schen Modells (2)

### Boole'sche Junktoren

- ◆ schwierige Anwendung Boole'scher Junktoren
- ◆ Verwechslung mit „und“, „oder“ und „nicht“

⇒ Impliziter weiterer Gegensatz:  
IR wird stärker von Laien benutzt als SQL



verschiedene Stufen der Relevanz durch Überführung  
Konjunktion in Disjunktion

Präsentation der Ergebnisse sortiert nach Relevanzstufen

Beispiel: Umwandlung von

„Korsika and Strand“ → „Korsika or Strand“

Ergebnis: zuerst die and-({d2}), dann restliche or-  
Dokumente ({d1, d3})

faceted query

zweistufiges Suchverfahren

1. Formulierung und Verfeinerung der Anfrage anhand benannter Anfragen und Ergebnisanzahl
2. Ergebnis zur finalen Anfrage anzeigen

## Beispiel:

- 1 Korsika liefert Q1: 1345
- 1 Q1 and Strand liefert Q2: 13
- 2 Anzeige Q2 liefert die 13 Ergebnisdokumente

Anwender sucht Dokumente über Korsika und Dokumente über Sardinien

“falsche” d.h. nicht intendierte Anfrage:

*Korsika and Sardinien*

liefert Dokumente, in denen beide Terme gemeinsam auftreten

“richtige” bzw. intendierte Anfrage:

*Korsika or Sardinien*

Verwendung besserer Junktorenbegriffe, etwa Ersetzen  
von

„and“ durch „all“

„or“ durch „any“

Erweiterung des Boole'schen Modells um Unschärfe (fuzzy)

Verallgemeinerung Boole'scher Junktoren

Unschärfe durch graduelle Zugehörigkeit von Dokumenten zu Termen

## Definition:

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

## Fuzzy-Mengen beim Information Retrieval

Universum ist Menge aller gespeicherten Dokumente

Term definiert Fuzzy-Menge

Zugehörigkeit (Fuzzy-Wert) des Dokuments  $d$  zu Term  $t$   
*durch Wert*

- ◆ **0 für keine Relevanz**
- ◆ **1 für maximale Relevanz**
- ◆ **Wert zwischen 0 und 1 für graduelle Relevanz**

$$\mu_t(d)$$

Universum umfasst 3 Dokumente

{d1, d2, d3}

Fuzzy-Mengen Korsika bzw. Strand drücken

Zugehörigkeit zu Term „Korsika“ bzw. „Strand“ aus

$$\text{Korsika} = \{ \langle d1; 0,1 \rangle, \langle d2; 0,6 \rangle, \langle d3; 1 \rangle \}$$

$$\text{Strand} = \{ \langle d1; 0,3 \rangle, \langle d2; 0,2 \rangle, \langle d3; 0,8 \rangle \}$$

$\mu$	d1	d2	d3
$\mu_{\text{Korsika}}$	0,1	0,6	1
$\mu_{\text{Strand}}$	0,3	0,2	0,8



jedes Dokument in jeder Fuzzy-Menge

→ übliche Mengenoperationen nicht anwendbar

Junktoren ermitteln neue Zugehörigkeitswerte

*and durch Min-Funktion*

*or durch Max-Funktion*

*not durch Subtraktion von 1:*

$$\begin{array}{lll} \text{Konjunktion } \mathbf{and}: & \mu_{A \cap B}(u) & = \min(\mu_A(u), \mu_B(u)) \\ \text{Disjunktion } \mathbf{or}: & \mu_{A \cup B}(u) & = \max(\mu_A(u), \mu_B(u)) \\ \text{Negation } \mathbf{not}: & \mu_{\bar{A}}(u) & = 1 - \mu_A(u) \end{array}$$

Korsika and Strand

Korsika or Strand

not Korsika

Anfrage	$\mu$	$d1$	$d2$	$d3$
1	$\mu_{\text{Korsika}}$	0,1	0,6	1
	$\mu_{\text{Strand}}$	0,3	0,2	0,8
2	$\mu_{\text{Korsika} \cap \text{Strand}}$	0,1	0,2	0,8
3	$\mu_{\overline{\text{Korsika}}}$	0,9	0,4	0

Überführung Anfrage in disjunktive Normalform  
jeder Suchterm induziert eine Fuzzy-Menge  
Anwendung entsprechender Fuzzy-Operationen auf  
Fuzzy-Mengen  
Ergebnis: Dokumente absteigend sortiert nach  
Zugehörigkeitsgrad ausgeben

Ergebnis umfasst alle Dokumente des Universums  
Begrenzung der Anzahl durch

- ◆ Schwellwerte für Zugehörigkeitswerte
- ◆ vorgegebene Anzahl von Ergebnisdokumenten

Beispiel:

Korsika and Strand

Schwellwert 0,5 liefert d3

Anzahl 2 liefert d2, d3

viele Möglichkeiten

Beispiel: Ansatz von Ogawa, Morita, Kobayashi mittels  
Term-zu-Term-Korrelationsmatrix

- ♦ ***Zeile entspricht Term  $i$  und Spalte entspricht Term  $j$***
- ♦  ***$n_{i,j}$  ist Anzahl Dokumente, welche Terme  $t_i, t_j$  enthalten***

## Dokument zu Term (2)

$n_i$  ist Anzahl Dokumente, welche Term  $t_i$  enthalten

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

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



Dokument d1: {Sardinien, Strand, Ferienwohnung}

Dokument d2: {Korsika, Strand, Ferienwohnung}

Dokument d3: {Korsika, Gebirge}

	$t_{Sardinien}$	$t_{Strand}$	$t_{Ferienw.}$	$t_{Korsika}$	$t_{Gebirge}$
$t_{Sardinien}$	1	0,5	0,5	0	0
$t_{Strand}$	0,5	1	1	0,333	0
$t_{Ferienw.}$	0,5	1	1	0,333	0
$t_{Korsika}$	0	0,333	0,333	1	0,5
$t_{Gebirge}$	0	0	0	0,5	1

Dokument d1: {**S**ardinien, **S**trand, **F**erienwohnung}

Dokument d2: {**K**orsika, **S**trand, **F**erienwohnung}

Dokument d3: {**K**orsika, **G**ebirge}

$$\mu_{\text{Sar}}(d_1) = 1 - \prod_{t_k \in \{\text{Sar, Str, Fer}\}} (1 - c_{i,k}) = 1 - (1 - 1)(1 - 0,5)(1 - 0,5) = 1$$

$$\mu_{\text{Kor}}(d_1) = 1 - \prod_{t_k \in \{\text{Sar, Str, Fer}\}} (1 - c_{i,k}) = 1 - (1 - 0)(1 - 0,33)(1 - 0,33) = 5/9$$

## Dokument zu Anfrageterm

Fuzzy-Modell für komplexe Ähnlichkeitsanfragen:  
Ähnlichkeitswert bzgl. atomarer Anfrage als  
Zugehörigkeitswert

anwendbar für jeden Medien-Typ

**Beispiel:** Suche nach roter Morgensonne

*Farbe = rot and Gestalt = Kreis*

# Vektorraummodelle

weit verbreitetes Retrieval-Modell

Dokumente als Vektoren eines Vektorraums

→ Überführung Retrieval-Problem in Gebiet der linearen Algebra

einsetzbar, wenn jedes Medien-Objekt darstellbar durch feste Anzahl von numerischen Merkmalswerten

Beispiel Bild-Retrieval: Vektorwerte etwa anhand der Farbverteilung

Beispiel Text-Retrieval:

- ◆ jeder Indexterm eine eigene Dimension
- ◆ Termgewicht (meist Häufigkeiten) als Vektorwert einer Dimension

Anfrage selbst als Vektor

Ähnlichkeit zwischen Anfrage  $q$  und Dokument  $d$  über deren Vektoren

- ◆ *Kosinusmaß (Kosinus des eingeschlossenen Winkels als Ähnlichkeitsmaß)*

$$sim_{cos}(d, q) = \frac{\langle d, q \rangle}{|d| * |q|}$$

- ◆ *Distanzmaß (Abstand zwischen Vektoren als Unähnlichkeitsmaß) z.B. Euklid'sche Distanz:*

$$dissim_{L_2}(d, q) = \sqrt{(d - q)^T (d - q)} = \sqrt{\sum_i (d[i] - q[i])^2}$$

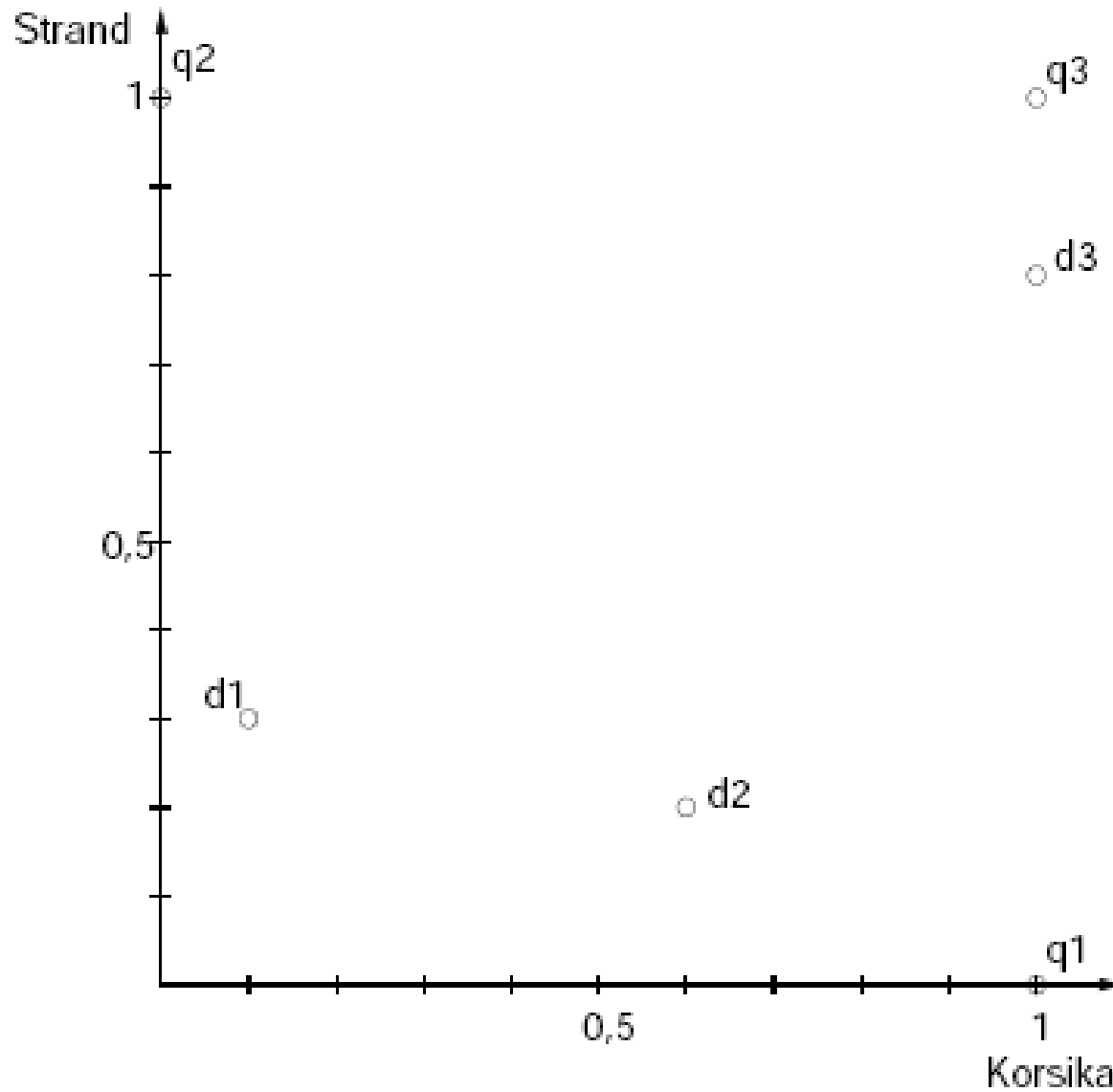
Dokumente:

<b>Dimension</b>	<i>d1</i>	<i>d2</i>	<i>d3</i>
Korsika	0,1	0,6	1
Strand	0,3	0,2	0,8

Anfragen:

<b>Dimension</b>	<i>q1</i>	<i>q2</i>	<i>q3</i>
Korsika	1	0	1
Strand	0	1	1

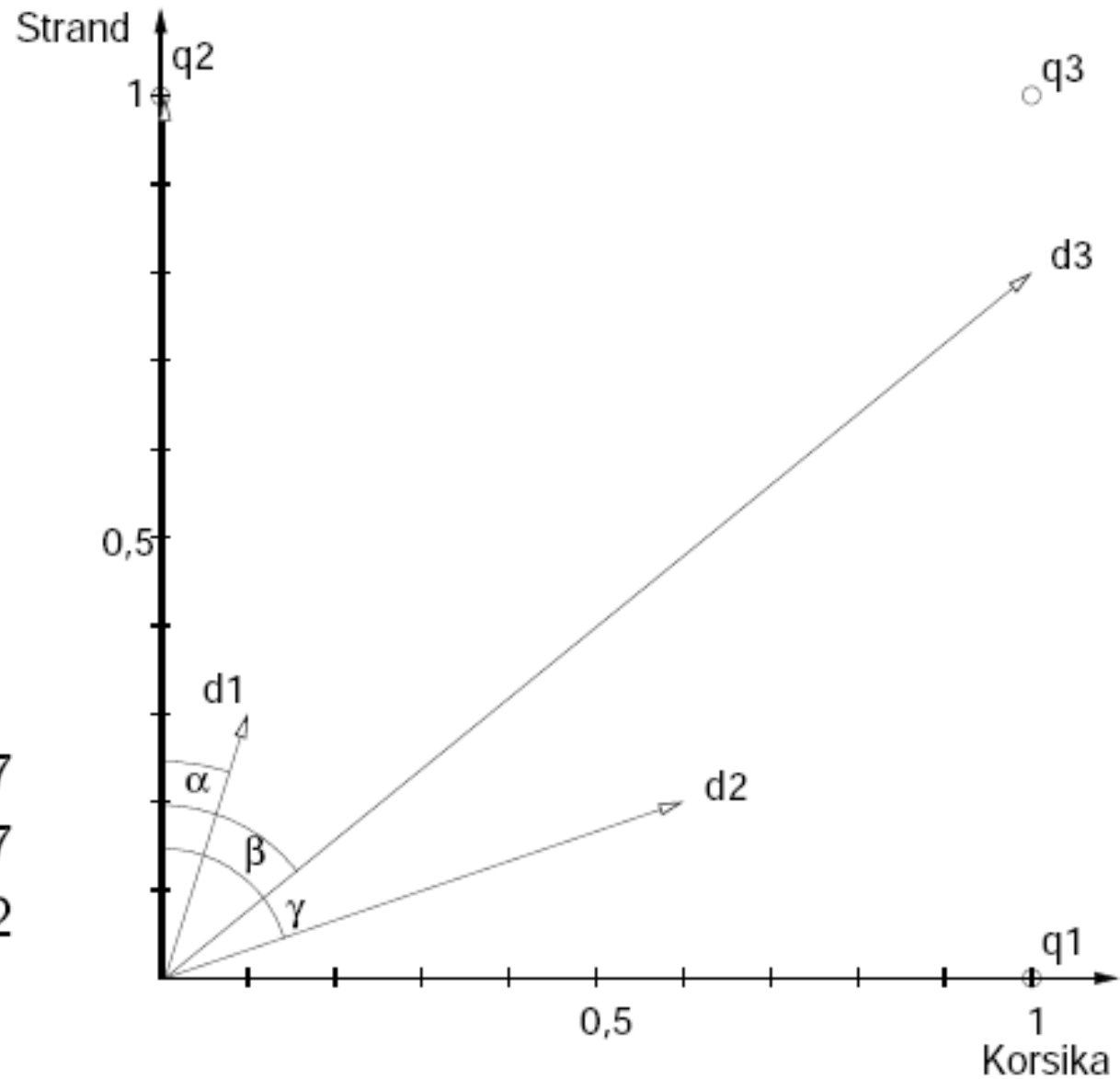




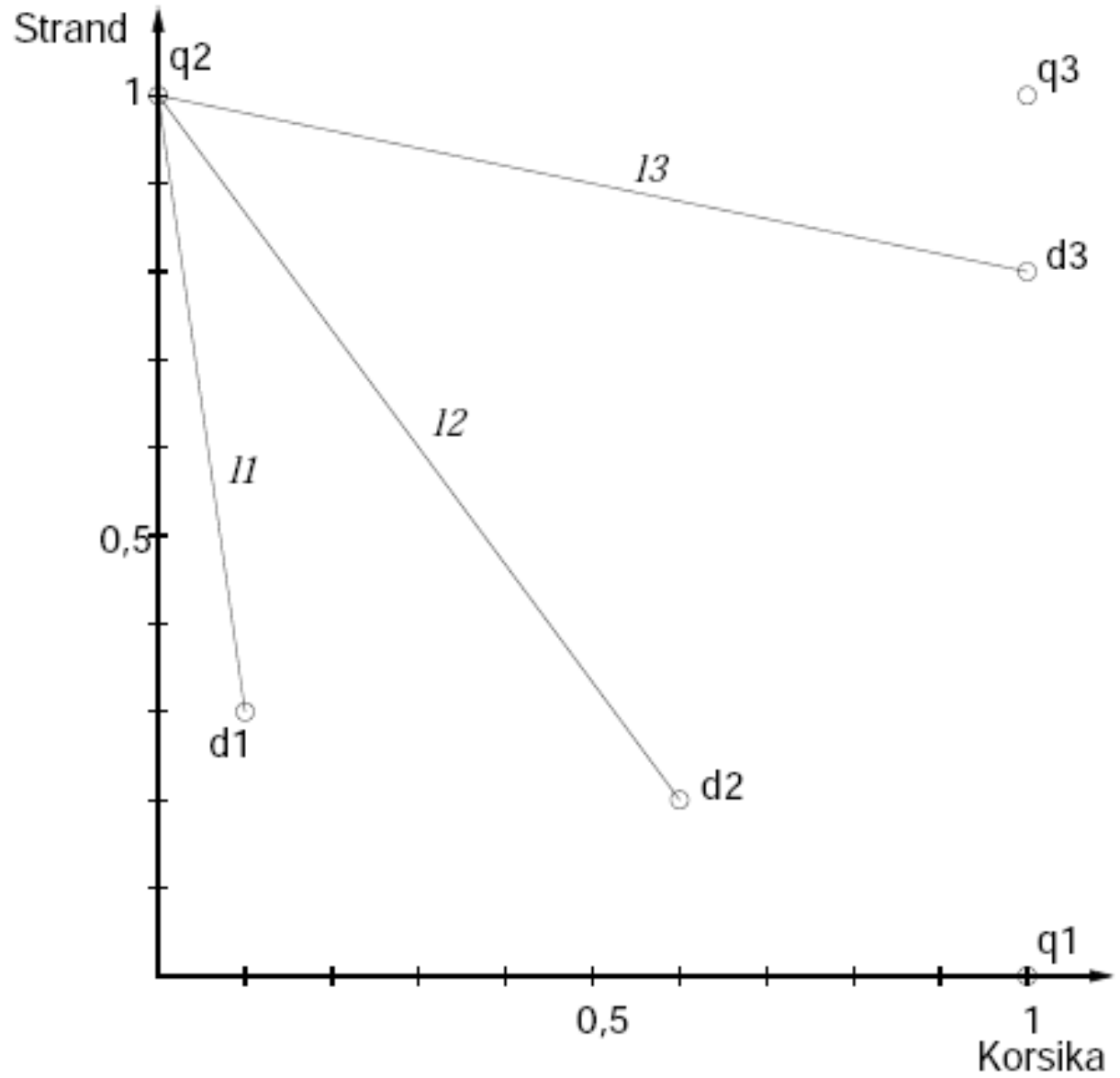
$sim_{cos}$	$d1$	$d2$	$d3$
$q1$	0,3162	0,9487	0,7809
$q2$	0,9487	0,3162	0,6247
$q3$	0,8944	0,8944	0,9939

# Eingeschlossene Winkel

$$\begin{aligned}\cos \alpha &= 0,9487 \\ \cos \beta &= 0,6247 \\ \cos \gamma &= 0,3162\end{aligned}$$



$dissim_{L_2}$	$d1$	$d2$	$d3$
$q1$	0,9487	0,4472	0,8
$q2$	0,7071	1	1,0198
$q3$	1,1402	0,8944	0,2



$$\begin{aligned} l1 &= 0,7071 \\ l2 &= 1 \\ l3 &= 1,0198 \end{aligned}$$

## Kosinusmaß und Euklid'sche Distanz

erzeugen unterschiedliche Ergebnisse (Sortierungen)

- ◆ Kosinusmaß erzeugt  $\langle d1, d3, d2 \rangle$
- ◆ Euklid'sche Distanz erzeugt  $\langle d1, d2, d3 \rangle$

Wahl der geeigneten Ähnlichkeitsfunktion abhängig von

- ◆ subjektivem Ähnlichkeitsempfinden
- ◆ Anwendungsszenario

Vektorraummodell sehr weit verbreitet  
setzt feste Anzahl von numerischen Merkmalswerten pro  
Dokument voraus

### Probleme:

- ◆ Merkmale als orthogonale Dimensionen aufgefasst (unrealistisch)  
→ Orthogonalisierung
- ◆ Problem bei hoher Anzahl von Merkmalswerten bzgl. Effektivität und Effizienz  
→ Dimensionsreduktion
- ◆ Anfrage ist Vektor, also keine Junktoren  
→ Boole'sche Junktoren durch Kombination mit dem Fuzzy-Modell