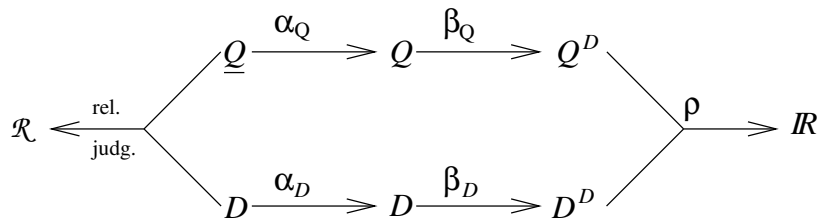


5. Nicht-Probabilistische Retrievalmodelle

Norbert Fuhr

Notationen



Q: Informationsbedürfnisse

Q: Frage-Repräsentationen

Q^D : Frage-Beschreibungen (formale Anfragen)

D: Dokumente

D: Dokument-Repräsentationen

D^D : Dokument-Beschreibungen (Dok.-Indexierungen)

\mathcal{R} : Relevanzskala

ρ : Retrievalfunktion

$T = \{t_1, \dots, t_n\}$: Indexierungsvokabular

$d_m^D: \vec{d}_m = (d_{m_1}, \dots, d_{m_n})$: Dokument-Beschreibung als Menge von Indexierungsgewichten

Überblick über die Modelle

	Bool.	Fuzzy	Vektor	Prob.	Cluster.
theoretische Basis: – boolesche Logik – Fuzzy-Logik – Vektoralgebra – Wahrsch.-Theorie	x	x	x	x	x
Bezug zur Retrievalqual.		(x)		x	
gewichtete Indexierung		x	x	x	x
gewichtete Frageterms		(x)	x	x	
Fragestruktur: – linear – boolesch	x	x	x (x)	x (x)	
Suchmodus: – Suchen – Browsen	x	x	x	x	x

Boolesches Retrieval

Historisch als erstes Retrievalmodell entwickelt und eingesetzt
(Dokument-Beschreibungen auf Magnetbändern!)

Dokumenten-Beschreibungen D^D :

ungewichtete Indexierung, d.h. $d_m^D = \vec{d}_m$ mit $d_{m_i} \in \{0, 1\}$ für
 $i = 1, \dots, n$

boolesches Retrieval liefert nur Zweiteilung der Dokumente in
"gefundene" ($q = 1$) und "nicht gefundene" ($q = 0$) Dokumente

Frage-Beschreibungen Q^D :

1. $t_i \in T \Rightarrow t_i \in Q^D$
2. $q_1, q_2 \in Q^D \Rightarrow q_1 \wedge q_2 \in Q^D$
3. $q_1, q_2 \in Q^D \Rightarrow q_1 \vee q_2 \in Q^D$
4. $q \in Q^D \Rightarrow \neg q \in Q^D$

Retrievalfunktion $\varrho(q, d_m)$:

1. $t_i \in T \Rightarrow \varrho(t_i, \vec{d}_m) = d_{m_i}$
2. $\varrho(q_1 \wedge q_2, \vec{d}_m) = \min(\varrho(q_1, \vec{d}_m), \varrho(q_2, \vec{d}_m))$
3. $\varrho(q_1 \vee q_2, \vec{d}_m) = \max(\varrho(q_1, \vec{d}_m), \varrho(q_2, \vec{d}_m))$
4. $\varrho(\neg q, \vec{d}_m) = 1 - \varrho(q, \vec{d}_m)$

Mächtigkeit der booleschen Anfragesprache:

jede beliebige Dokumentenmenge kann selektiert werden
(Voraussetzung: alle Dokumente besitzen unterschiedliche Indexierungen)

Konstruktion der booleschen Frageformulierung q_k zu einer vorgegebenen Dokumentenmenge D_k :

$$\begin{aligned}d_m^Q &= x_{m_1} \wedge \dots \wedge x_{m_n} \text{ mit} \\x_{m_i} &= \begin{cases} t_i & \text{falls } d_{m_i} = 1 \\ \neg t_i & \text{sonst} \end{cases} \\q_k &= \bigvee_{d_j \in D_k} d_j^Q\end{aligned}$$

Beispiel-Recherche

“The side effects of drugs on memory or cognitive abilities, not related to aging”

1. 19248 DRUGS
2. 2412 DRUGS in TI
3. 2560 AGING
4. 19119 DRUG not AGING
5. 2349 #2 and #4
6. 9305 MEMORY
7. 6 #5 and (DRUG near4 MEMORY)
8. 22091 COGNITIVE
9. 16 #5 and (DRUG near4 COGNITIVE)
10. 22 #7 or #9
11. 2023 SIDE-EFFECTS-DRUG in DE
12. 0 #11 and #10

Nachteile des booleschen Retrieval

1. Größe der Antwortmenge ist schwierig zu kontrollieren
2. Keine Ordnung der Antwortmenge nach mehr oder weniger relevanten Dokumenten
3. Keine Möglichkeit zur Gewichtung von Fragetermen oder gewichteter Indexierung
4. Trennung in gefundene und nicht gefundene Dokumente zu streng:
Zu $q = t_1 \wedge t_2 \wedge t_3$ werden Dokumente mit zwei gefundenen Termen genauso zurückgewiesen wie solche mit 0
Analog für $q = t_1 \vee t_2 \vee t_3$ keine Unterteilung der gefundenen Dokumente
5. Erstellung der Frageformulierung sehr umständlich
6. schlechte Retrievalqualität

Fuzzy-Retrieval

Teilweise Überwindung der Nachteile des booleschen Retrieval

Dokumenten-Beschreibungen:

Erweiterung auf gewichtete Indexierung, d.h. $d_{m_i} \in [0, 1]$

Frage-Beschreibungen, Retrievalfunktion:

wie beim booleschen Retrieval

Retrievalfunktion liefert jetzt Werte $\varrho(q_k^D, \vec{d}_m) \in [0, 1]$

→ Ranking der Antwortmenge

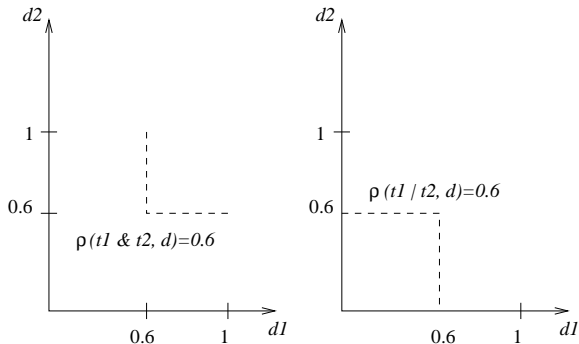
aber: Definition der Retrievalfunktion umstritten

$$T = \{t_1, t_2\}$$

$$q = t_1 \wedge t_2$$

$$\vec{d}_1 = (0.6, 0.6) \quad , \quad \vec{d}_2 = (0.59, 0.99)$$

$$\varrho(q, \vec{d}_1) = 0.6 \quad , \quad \varrho(q, \vec{d}_2) = 0.59$$



Andere Definitionen der Fuzzy-Operatoren (T-Normen)
überwinden Nachteile der Standard-Definition,
aber verletzen Gesetze der Booleschen Algebra:
(z.B. $\rho(((t_1 \vee t_2) \wedge t_3), d) \neq \rho(((t_1 \wedge t_3) \vee (t_2 \wedge t_3)), d)$)

Kollektion	MEDLARS	ISI	INSPEC	CACM
#Dok.	1033	1460	12684	3204
#Fragen	30	35	77	52
Bool.	0.2065	–	0.1159	–
Fuzzy	0.2368	0.1000	0.1314	0.1551
Vektor	0.5473	0.1569	0.2325	0.3027

Experimenteller Vergleich von Booleschem Retrieval,
Fuzzy-Retrieval und Vektorraummodell

Beurteilung des Fuzzy-Retrieval

- + Generalisierung des booleschen Retrieval für gewichtete Indexierung → Ranking
- keine Fragetermgewichtung
- schlechte Retrievalqualität
- Erstellung der Frageformulierung sehr umständlich

Das Vektorraummodell

Definition

zuerst entstanden im Rahmen der Arbeiten zu SMART
(experimentelles Retrievalsystem von G. Salton und Mitarbeitern
(Harvard/Cornell), seit 1961)

Dokumente und Fragen als Punkte in einem orthonormalen
Vektorraum, der durch die Terme aufgespannt wird

orthonormaler Vektorraum:

- ▶ alle Term-Vektoren orthogonal (und damit auch linear unabhängig)
- ▶ alle Term-Vektoren normiert

Dokument-Beschreibung: ähnlich wie Fuzzy-Retrieval

$$d_m^D = \vec{d}_m \text{ mit } d_{m_i} \in \mathbf{R} \text{ für } i = 1, \dots, n$$

Frage-Beschreibung:

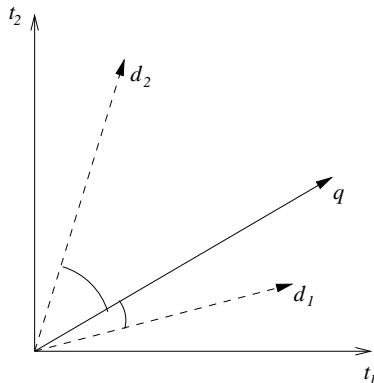
$$q_k^Q = \vec{q}_k \text{ mit } q_{k_i} \in \mathbf{R} \text{ für } i = 1, \dots, n$$

Retrievalfunktion

Vektor-Ähnlichkeitsmaße, z.B. Cosinus

Meistens: Skalarprodukt

$$\varrho(\vec{q}_k, \vec{d}_m) = \vec{q}_k \cdot \vec{d}_m$$



Beispiel-Frage:

“side effects of drugs on memory and cognitive abilities”

t_i	q_{k_i}	d_{1_i}	d_{2_i}	d_{3_i}	d_{4_i}
side effect	2	1	0.5	1	1
drugs	2	1	1	1	1
memory	1	1		1	
cognitive ability	1		1	1	0.5
Retrievalgewicht		5	4	6	4.5

Coordination Level Match

Vereinfachung des Vektorraummodells:

nur binäre Frage- und Dokumenttermgewichtung

Dokument-Beschreibung: ähnlich wie Boolesches Retrieval

$d_m^D = \vec{d}_m$ mit $d_{m_i} \in \{0, 1\}$ für $i = 1, \dots, n$

Frage-Beschreibung:

$q_k^Q = \vec{q}_k$ mit $q_{k_i} \in \{0, 1\}$ für $i = 1, \dots, n$

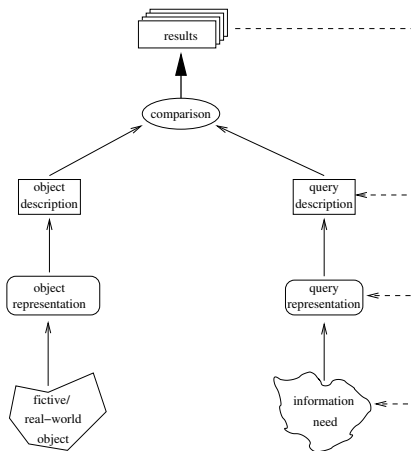
Retrievalfunktion:

Skalarprodukt

$$\rho(\vec{q}_k, \vec{d}_m) = \vec{q}_k \cdot \vec{d}_m = |q_k^T \cap d_m^T|$$

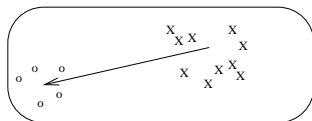
Relevance Feedback

iteratives Retrieval:



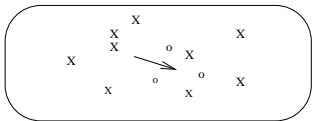
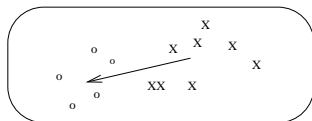
Relevance Feedback im VRM

Ziel: Modifikation des Fragevektors



o: relevant

X: irrelevant



Bestimmung des optimalen Fragevektors

D^R : relevante Dokumente

D^N : irrelevante Dokumente

Idee:

wähle Fragevektor \vec{q} so, dass Differenz der RSVs zwischen relevanten und irrelevanten Dokumenten maximal wird:

$$\sum_{(d_k, d_l) \in D^R \times D^N} \vec{q} \vec{d}_k - \vec{q} \vec{d}_l \stackrel{!}{=} \max$$

mit der Nebenbedingung $\sum_{i=1}^n q_i^2 = c$

Extremwertproblem mit Randbedingung

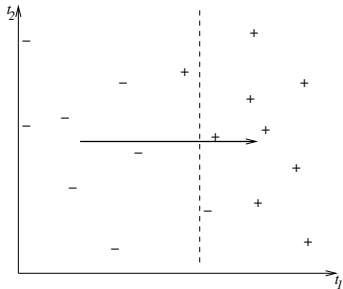
→ Lagrange-Multiplikator einsetzen

$$\begin{aligned}
F &= \lambda \left(\sum_{i=1}^n q_i^2 - c \right) + \sum_{(d_k, d_l) \in D^R \times D^N} \sum_{i=1}^n q_i d_{k_i} - q_i d_{l_i} \\
\frac{\partial F}{\partial q_i} &= 2\lambda q_i + \sum_{(d_k, d_l) \in D^R \times D^N} d_{k_i} - d_{l_i} \stackrel{!}{=} 0 \\
q_i &= -\frac{1}{2\lambda} \sum_{(d_k, d_l) \in D^R \times D^N} d_{k_i} - d_{l_i} \\
\vec{q} &= -\frac{1}{2\lambda} \sum_{(d_k, d_l) \in D^R \times D^N} \vec{d}_k - \vec{d}_l \\
&= -\frac{1}{2\lambda} \left(|D^N| \sum_{d_k \in D^R} \vec{d}_k - |D^R| \sum_{d_l \in D^N} \vec{d}_l \right) \\
&= -\frac{|D^N| |D^R|}{2\lambda} \left(\frac{1}{|D^R|} \sum_{d_k \in D^R} \vec{d}_k - \frac{1}{|D^N|} \sum_{d_l \in D^N} \vec{d}_l \right)
\end{aligned}$$

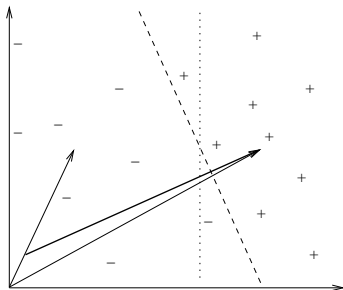
wähle c so, dass $|D^N| |D^R| / 2\lambda = -1$:

$$\vec{q} = \frac{1}{|D^R|} \sum_{d_k \in D^R} \vec{d}_k - \frac{1}{|D^N|} \sum_{d_l \in D^N} \vec{d}_l$$

$\hat{=}$ Verbindungsvektor der Zentroiden der relevanten / irrelevanten Dokumente



unterschiedliche Gewichtung positiver und negativer Beispiele:



Rocchio-Algorithmus

- ▶ unterschiedliche Gewichtung positiver und negativer Beispiele
- ▶ Berücksichtigung der ursprünglichen Anfrage

$$\vec{q}_k' = \vec{q}_k + \alpha \frac{1}{|D_k^R|} \sum_{d_j \in D_k^R} \vec{d}_j - \beta \frac{1}{|D_k^N|} \sum_{d_j \in D_k^N} \vec{d}_j$$

α, β — positive Konstanten, heuristisch festzulegen (z.B. $\alpha = 0.75, \beta = 0.25$)

Vorgehensweise:

1. Retrieval mit Fragevektor \vec{q}_k vom Benutzer
2. Relevanzbeurteilung der obersten Dokumente der Rangordnung
3. Berechnung eines verbesserten Fragevektors \vec{q}_k' aufgrund der Feedback-Daten
4. Retrieval mit dem verbesserten Vektor
5. Evtl. Wiederholung der Schritte 2-4

Dokumenten-Indexierung

Vektorraum-Modell liefert keine Aussagen darüber, wie die Dokumenten-Indexierung zu berechnen ist!

(Dokumenten-)Indexierung im Vektorraummodell:
heuristische Formeln zur Berechnung der Indexierungsgewichte
zugrundeliegende Dokumenten-Repräsentation: Multi-Menge (Bag)
von Termen

Heuristiken:

Indexierungsgewicht umso höher, je ...

- ▶ seltener der Term in der Kollektion
- ▶ häufiger der Term im Dokument
- ▶ kürzer das Dokument

- d_m^T Menge der in d_m vorkommenden Terms
- l_m Dokumentlänge (# laufende Wörter in d_m)
- al durchschnittliche Dokumentlänge in \underline{D}
- tf_{mi} : Vorkommenshäufigkeit (Vkh) von t_i in d_m .
- n_i : # Dokumente, in denen t_i vorkommt.
- $|\underline{D}|$: # Dokumente in der Kollektion

inverse Dokumenthäufigkeit (idf):

$$idf_i = \frac{\log \frac{|\underline{D}|}{n_i}}{|\underline{D}| + 1}$$

normalisierte Vorkommenshäufigkeit:

$$ntf_i = \frac{tf_{mi}}{tf_{mi} + 0.5 + 1.5 \frac{l_m}{al}}$$

Indexierungsgewicht tfidf:

$$w_{mi} = ntf_i \cdot idf_i$$

Kollektion	CACM	CISI	CRAN	INSPEC	MED
Coord.	0.185	0.103	0.241	0.094	0.413
SMART	0.363	0.219	0.384	0.263	0.562

Binäre Gewichte (Coordination Level Match) vs.
SMART-Gewichtung von Fragen und Dokumenten
(aus Salton/Buckley 88)

Beurteilung des Vektorraummodells

- + einfaches Modell, insbes. für den Benutzer
- + unmittelbar anwendbar auf neue Kollektionen
- + gute Retrievalqualität
- sehr viele heuristische Komponenten
- kein Bezug zur Retrievalqualität
(Optimalität von Relevance Feedback?)
- Dokumentrepräsentation kann schlecht erweitert werden