

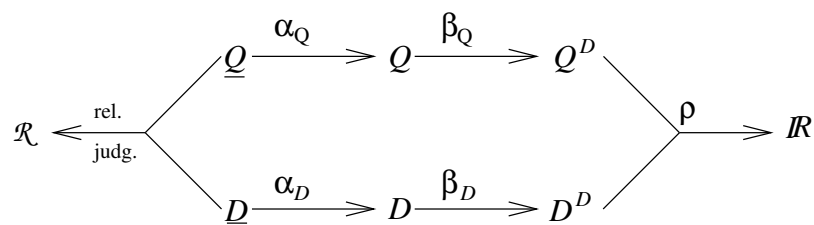
Internet-Suchmaschinen Probabilistische Retrievalmodelle

Norbert Fuhr

Notationen

1 / 41

Notationen



- | | |
|---|----------------------------------|
| $q \in \underline{Q}$ Anfrage/Info-bed. | $d_m \in \underline{D}$ Dokument |
| $q \in Q$ Anfragerepräs. | $d \in D$ Dokumentrepräs. |
| $q^D \in Q^D$ Anfragebeschr. | $d_m^D \in D^D$ Dokumentbeschr. |
| \mathcal{R} Relevanzskala | ρ Retrievalfunktion |
| | R Retrievalwert |

Binary-Independence-Retrieval-Modell

Retrievalfunktionen für binäre Indexierung
Probabilistische Grundlagen des BIR-Modells
Anwendung des BIR-Modells

3 / 41

Retrievalfunktionen für binäre Indexierung

repräsentiere Anfragen und Dokumente als Mengen von Termen
 $T = \{t_1, \dots, t_n\}$ Menge der Terme in einer Kollektion

$q \in Q$: Anfragerrepräsentation

q^T : Menge von Fragetermen

$d_m \in D$: Dokumentrepräsentation

d_m^T : Menge von Dokumenttermen

einfache Retrievalfunktion: **Coordination level match**

$$O_{COORD}(q, d_m) = |q^T \cap d_m^T|$$

Binary-Independence-Retrieval-Modell (BIR):

weise Fragetermen Gewichte zu

$$O_{BIR}(q, d_m) = \sum_{t_i \in q^T \cap d_m^T} c_i$$

5/41

Probabilistische Grundlagen des BIR-Modells

Grundlegende mathematische Techniken zur Herleitung der probabilistischen Retrievalmodelle:

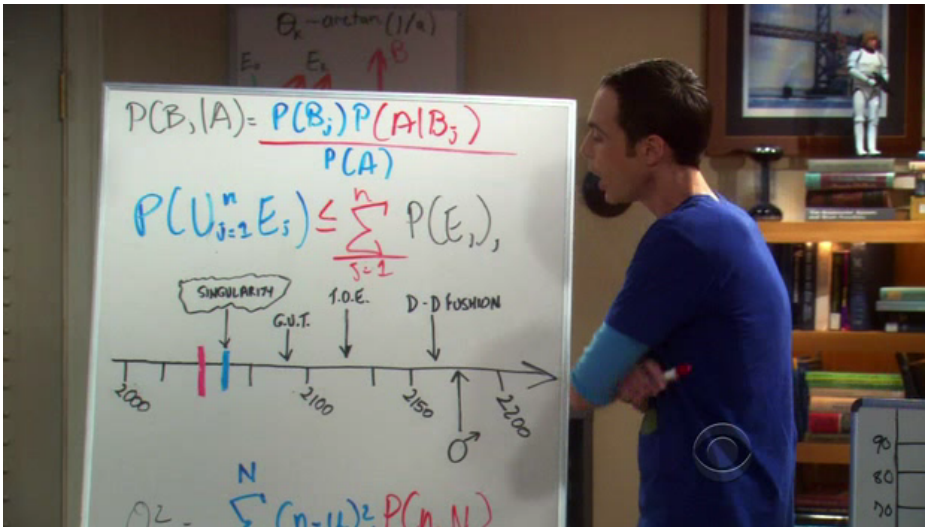
1. Benutzung von Chancen statt Wahrscheinlichkeiten, wobei

$$O(y) = \frac{P(y)}{P(\bar{y})} = \frac{P(y)}{1 - P(y)}$$

2. Anwendung des Bayes'schen Theorems:

$$P(a|b) = \frac{P(a, b)}{P(b)} = \frac{P(b|a) \cdot P(a)}{P(b)}$$

6/41



Szene aus „The Big Bang Theory“ S04E02

Herleitung des BIR-Modells

Abschätzung von $O(R|d_m^T)$

= Chance, dass ein Dokument mit einer Menge von Termen d_m^T relevant zur Anfrage q ist

Repräsentation des Dokumentes d_m als binären Vektor

$$\vec{x} = (x_1, \dots, x_n) \quad \text{mit} \quad x_i = \begin{cases} 1, & \text{falls } t_i \in d_m^T \\ 0, & \text{sonst} \end{cases}$$

$$O(R|d_m^T) = O(R|\vec{x}) = \frac{P(R|\vec{x})}{P(\bar{R}|\vec{x})}$$

8/41

Anwenden des Bayes'schen Theorems

$$O(R|\vec{x}) = \frac{P(R|\vec{x})}{P(\bar{R}|\vec{x})} = \frac{P(R)}{P(\bar{R})} \cdot \frac{P(\vec{x}|R)}{P(\vec{x}|\bar{R})} \cdot \frac{P(\vec{x})}{P(\vec{x})}$$

$P(R)$ W., dass ein arbiträres Dokument relevant ist zur Anfrage

$P(\vec{x}|R)$ W., dass ein arbiträres, relevantes Dokument den Termvektor \vec{x} besitzt

$P(\vec{x}|\bar{R})$ W., dass ein arbiträres, nicht-relevantes Dokument den Termvektor \vec{x} besitzt

Beispiel

d_m	$r(d_m)$	x_1	x_2	$P(R \vec{x})$
d_1	R	1	1	0.80
d_2	R	1	1	
d_3	R	1	1	
d_4	R	1	1	
d_5	N	1	1	
d_6	R	1	0	0.67
d_7	R	1	0	
d_8	R	1	0	
d_9	R	1	0	
d_{10}	N	1	0	
d_{11}	N	1	0	

d_m	$r(d_m)$	x_1	x_2	$P(R \vec{x})$
d_{12}	R	0	1	0.50
d_{13}	R	0	1	
d_{14}	R	0	1	
d_{15}	N	0	1	
d_{16}	N	0	1	
d_{17}	N	0	1	
d_{18}	R	0	0	0.33
d_{19}	N	0	0	
d_{20}	N	0	0	

$$P(R) = \frac{12}{20} \quad P(1,1|R) = \frac{4}{12} \quad P(1,1|\bar{R}) = \frac{1}{8}$$

9/41

10/41

Annahme der "Linked dependence":

$$\frac{P(\vec{x}|R)}{P(\vec{x}|\bar{R})} \approx \prod_{i=1}^n \frac{P(x_i|R)}{P(x_i|\bar{R})}$$

$$O(R|\vec{x}) = \frac{P(R)}{P(\bar{R})} \cdot \frac{P(\vec{x}|R)}{P(\vec{x}|\bar{R})} \approx O(R) \prod_{i=1}^n \frac{P(x_i|R)}{P(x_i|\bar{R})}$$

Aufteilen nach Vorkommen/Fehlen von Termen im aktuellen Dokument:

$$O(R|\vec{x}) = O(R) \prod_{x_i=1} \frac{P(x_i=1|R)}{P(x_i=1|\bar{R})} \cdot \prod_{x_i=0} \frac{P(x_i=0|R)}{P(x_i=0|\bar{R})}$$

$p_i = P(x_i=1|R)$ Wahrscheinlichkeit, dass t_i in einem arbiträren relevanten Dokument vorkommt

$s_i = P(x_i=1|\bar{R})$ Wahrscheinlichkeit, dass t_i in einem arbiträren nicht-relevanten Dokument vorkommt

Annahme, dass $p_i = s_i$ für alle $t_i \notin q^T$

$$\begin{aligned} O(R|d_m^T) &= O(R) \prod_{t_i \in d_m^T \cap q^T} \frac{p_i}{s_i} \cdot \prod_{t_i \in q^T \setminus d_m^T} \frac{1-p_i}{1-s_i} \\ &= O(R) \prod_{t_i \in d_m^T \cap q^T} \frac{p_i}{s_i} \cdot \prod_{t_i \in d_m^T \cap q^T} \frac{1-s_i}{1-p_i} \\ &\quad \cdot \prod_{t_i \in d_m^T \cap q^T} \frac{1-p_i}{1-s_i} \cdot \prod_{t_i \in q^T \setminus d_m^T} \frac{1-p_i}{1-s_i} \\ &= O(R) \prod_{t_i \in d_m^T \cap q^T} \frac{p_i(1-s_i)}{s_i(1-p_i)} \cdot \prod_{t_i \in q^T} \frac{1-p_i}{1-s_i} \end{aligned} \quad (1)$$

Nur das erste Produkt ist bezüglich einer gegebenen Anfrage q für unterschiedliche Dokumente *nicht* konstant \rightarrow

Betrachte daher nur dieses Produkt für das Ranking

11/41

12/41

Anwendung des BIR-Modells

Parameterabschätzung für s_i

$$O(R|d_m^T) = O(R) \prod_{t_i \in d_m^T \cap q^T} \frac{p_i(1-s_i)}{s_i(1-p_i)} \cdot \prod_{t_i \in q^T} \frac{1-p_i}{1-s_i}$$

Übergang zum Logarithmus (ordnungserhaltend):

$$c_i = \log \frac{p_i(1-s_i)}{s_i(1-p_i)}$$

Retrievalfunktion:

$$q_{BIR}(q, d_m) = \sum_{t_i \in d_m^T \cap q^T} c_i$$

$$s_i = P(x_i=1|\bar{R}):$$

(Wahrscheinlichkeit, dass t_i in einem arbiträren nicht-relevanten Dokument vorkommt)

Annahme:

Anzahl der nicht-relevanten Dokumente \approx Größe der Kollektion

N – Kollektionsgröße

n_i – # Dokumente mit dem Term t_i

$$s_i = \frac{n_i}{N}$$

13/41

14/41

Parameterabschätzung für p_i

$$p_i = P(x_i=1|R):$$

(Wahrscheinlichkeit, dass t_i in einem arbiträren relevanten Dokument vorkommt)

1. benutze globalen Wert p für alle p_i s

→ Termgewichtung nach inverser Dokumentenhäufigkeit (IDF)

$$\begin{aligned} c_i &= \log \frac{p}{1-p} + \log \frac{1-s_i}{s_i} \\ &= c_p + \log \frac{N-n_i}{n_i} \end{aligned}$$

$$q_{IDF}(q, d_m) = \sum_{t_i \in q^T \cap d_m^T} (c_p + \log \frac{N-n_i}{n_i})$$

oft benutzt: $p = 0.5 \rightarrow c_p = 0$

2. Relevance Feedback:

initiale Rangordnung nach IDF-Formel

präsentiere höchstgerankte Dokumente dem Benutzer
(etwa 10...20)

Benutzer gibt binäre Relevanzurteile ab: relevant/nicht-relevant

r : # als relevant beurteilte Dokumente zur Anfrage q

r_i : # relevante Dokumente mit dem Term t_i

$$p_i = P(t_i|R) \approx \frac{r_i}{r}$$

verbesserte Abschätzungen:

$$p_i \approx \frac{r_i + 0.5}{r + 1}$$

15/41

16/41

Beispiel für BIR

d_m	$r(d_m)$	x_1	x_2	$P(R \vec{x})$	BIR
d_1	R	1	1	0.80	0.76
d_2	R	1	1		
d_3	R	1	1		
d_4	R	1	1		
d_5	N	1	1		
d_6	R	1	0	0.67	0.69
d_7	R	1	0		
d_8	R	1	0		
d_9	R	1	0		
d_{10}	N	1	0		
d_{11}	N	1	0		

d_m	$r(d_m)$	x_1	x_2	$P(R \vec{x})$	BIR
d_{12}	R	0	1	0.50	0.48
d_{13}	R	0	1		
d_{14}	R	0	1		
d_{15}	N	0	1		
d_{16}	N	0	1		
d_{17}	N	0	1	0.33	0.40
d_{18}	R	0	0		
d_{19}	N	0	0		
d_{20}	N	0	0		

BIR Example

For the example collection above, compute the values of $O(R|d_m^T)$ via eqn. 1, estimating the parameters directly as relative frequencies. ▶

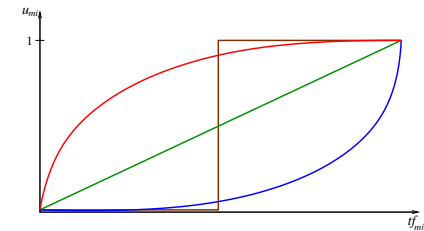
17 / 41

18 / 41

BM25

BM25

heuristische Erweiterung des BIR-Modells von binärer auf gewichtete Indexierung (Berücksichtigung der Vorkommenshäufigkeit tf)



20 / 41

Übergang zu gewichteter Indexierung

- l_m Dokumentlänge (# laufende Wörter in d_m)
- al durchschnittliche Dokumentlänge in \underline{D}
- tf_{mi} : Vorkommenshäufigkeit (Vkh) von t_i in d_m .
- b Gewichtung der Längennormalisierung, $0 \leq b \leq 1$
- k Gewichtung der Vorkommenshäufigkeit

Längennormalisierung: $B = \left((1 - b) + b \frac{l_m}{al} \right)$

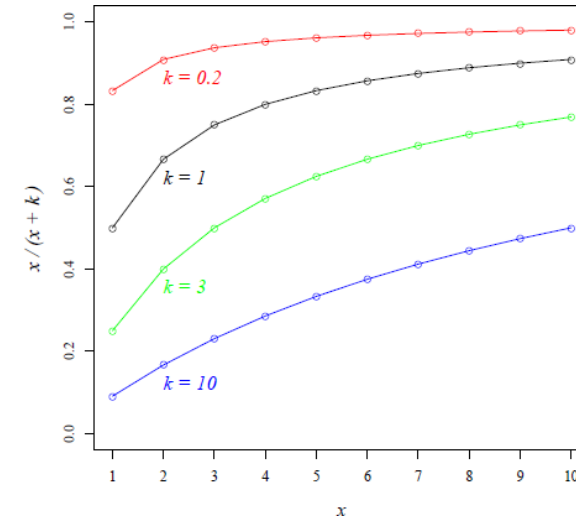
normalisierte Vorkommenshäufigkeit: $ntf_{mi} = tf_{mi} / B$

BM25-Gewicht:
$$u_{mi} = \frac{ntf_{mi}}{k + ntf_{mi}}$$

$$= \frac{tf_{mi}}{k \left((1 - b) + b \frac{l_m}{al} \right) + tf_{mi}}$$

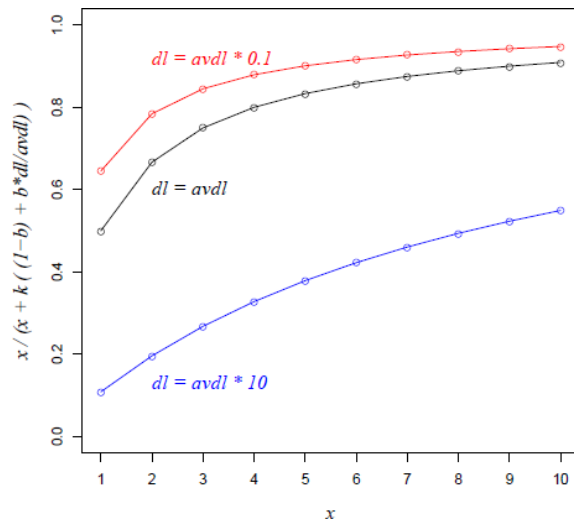
21 / 41

Einfluss von k



22 / 41

Einfluss von B



23 / 41

BM25-Retrievalfunktion

$$q_{BM25}(q, d_m) = \sum_{t_i \in d_m^T \cap q^T} u_{mi} \cdot c_i$$

$$= \sum_{t_i \in d_m^T \cap q^T} \frac{tf_{mi}}{k \left((1 - b) + b \frac{l_m}{al} \right) + tf_{mi}} \log \frac{p_i(1 - s_i)}{s_i(1 - p_i)}$$

24 / 41

Statistische Sprachmodelle

Sprachmodell von Zhai und Lafferty
 Ähnlichkeit von Wahrscheinlichkeitsverteilungen

Sprachmodell von Zhai und Lafferty

W., dass Anfrage vom Sprachmodell des Dokumentes generiert wurde:

$$\begin{aligned}
 P(q|d_m) &\approx \prod_{t_i \subseteq q^T} P(t_i|\theta_{d_m}) \\
 &= \prod_{t_i \in q^T \cap d_m^T} P_s(t_i|d_m) \prod_{t_i \in q^T - d_m^T} P_u(t_i|d_m) \\
 &= \prod_{t_i \in q^T \cap d_m^T} \frac{P_s(t_i|d_m)}{P_u(t_i|d_m)} \prod_{t_i \in q^T} P_u(t_i|d_m)
 \end{aligned}$$

$P_s(t_i|d_m)$ W. dass das Dokument über t_i ist, falls $t_i \in d^T$

$P_u(t_i|d_m)$ W. dass das Dokument über t_i ist, falls $t_i \notin d^T$

$P(t_i|\theta_{d_m}) = P_s(t_i|d)$, falls $t_i \in d^T$, $= P_u(t_i|d)$ sonst

Statistische Sprachmodelle

Nachteil bisher vorgestellter Modelle:
 keine theoretisch fundierte Berechnung der Indexierungsgewichte

Statistische Sprachmodelle:

- ▶ betrachten Sprache (Folge von Wörtern) als statistischen Prozess
- ▶ Sprachmodell θ ist definiert als Wahrscheinlichkeitsverteilung

$$\theta = \{(t_i, P(t_i|\theta)) | t_i \in T\} \quad \text{mit} \quad \sum_{t_i \in T} P(t_i|\theta) = 1$$

- ▶ Wahrscheinlichkeit für einen Dokumenttext $d = t_1 t_2 t_3 \dots t_l$:
 $P(d|\theta) = \prod_{j=1}^l P(t_j|\theta)$
- ▶ Retrievalfunktion: betrachte Wahrscheinlichkeit, dass Frage und Dokument vom selben Sprachmodell generiert wurden

26 / 41

Parameterschätzung:

Schätzung von $P_s(t_i|d_m)$: Problem aufgrund spärlicher Daten

L Anzahl Token der Kollektion

tf_{im} Vorkommenshäufigkeit von t_i in d_m

l_m Dokumentlänge (Anzahl Token) von d_m

cf_i Kollektionshäufigkeit von t_i (# Vorkommen)

$$P_{avg}(t_i) = \frac{cf_i}{L} \quad P_{ML}(t_i|d_m) = \frac{tf_{im}}{l_m}$$

schätze

$$P_s(t_i|d_m) = (1 - \lambda)P_{ML}(t_i|d_m) + \lambda P_{avg}(t_i)$$

$$P_u(t_i|d_m) = \alpha_m P_{avg}(t_i)$$

$0 \leq \lambda \leq 1$: Glättungsfaktor (Jelinek-Mercer)

$$\alpha_m = \frac{1 - \sum_{t_i \in q^T \cap d_m^T} P_{avg}(t_i)}{1 - \sum_{t_i \in q^T \cap d_m^T} P_{ML}(t_i|d_m)}$$

Exercise for the Zhai-Lafferty Model

Given the following collection of documents:

- ▶ $d_1 = (t_1, t_1, t_1, t_2)$
- ▶ $d_2 = (t_1, t_1, t_3, t_3)$
- ▶ $d_3 = (t_1, t_2, t_2)$
- ▶ $d_4 = (t_2)$

Now consider the query $q = (t_1, t_2)$.

Compute the language model probabilities according to the Zhai-Lafferty model.

Let $\lambda = 0.5$ and assume $\alpha_d = 1$ ▶

Das Probability-Ranking-Principle (PRP)

Entscheidungstheoretische Rechtfertigung des PRP
Rechtfertigung in Bezug auf Qualitätsmaße

Ähnlichkeit von Wahrscheinlichkeitsverteilungen

alternative Retrievalfunktion: **Kullback-Leibler Divergence** misst die Unähnlichkeit zweier statistischer Sprachmodelle

- ▶ Dokument-Sprachmodell θ_d (wie oben)
- ▶ Anfrage-Sprachmodell θ_q (z.B. als $P_{ML}(t|q)$)

Idee: messe relative Information

Information eines Terms: $-\log P(t|\theta)$

Differenz der Information: $\log P(t|\theta_q) - \log P(t|\theta_d) = \log \frac{P(t|\theta_q)}{P(t|\theta_d)}$

anschließend Gewichtung entsprechend der relativen Häufigkeit des Terms:

$$D(\theta_q||\theta_d) = \sum_{t_i \in q^T} P(t_i|\theta_q) \log \frac{P(t_i|\theta_q)}{P(t_i|\theta_d)}$$

29 / 41

30 / 41

Das Probability-Ranking-Principle (PRP)

Perfektes Retrieval:

ordne alle relevanten Dokumenten vor allen nicht-relevanten an
bezieht sich auf die Retrievalobjekte selbst, und ist nur bei vollständiger Relevanzbeurteilung der Kollektion möglich

Optimales Retrieval:

bezieht sich auf die Repräsentationen (wie jedes IR-System)

Probability Ranking Principle (PRP)

definiert optimales Retrieval für probabilistische Modelle:

ordne die Dokumente nach der absteigenden Wahrscheinlichkeit der Relevanz

32 / 41

Entscheidungstheoretische Rechtfertigung des PRP

\bar{C} : Kosten für Retrieval eines nicht-relevanten Dokumentes

C : Kosten für Retrieval eines relevanten Dokumentes

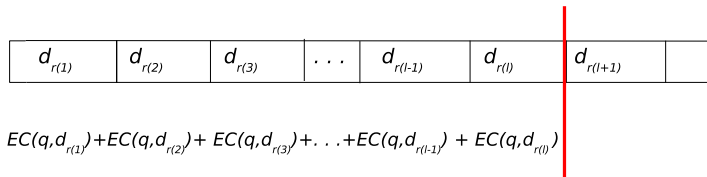
erwartete Kosten für das Retrieval eines Dokuments d_j :

$$EC(q, d_j) = C \cdot P(R|q, d_j) + \bar{C}(1 - P(R|q, d_j))$$

Gesamtkosten für das Retrieval:

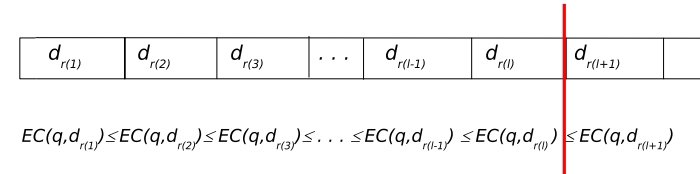
(angenommen, der Benutzer betrachtet die ersten l Dokumente, wobei l nicht im Voraus bekannt ist)

$r(i)$: Ranking-Funktion, bestimmt den Index des Dokuments für den Rang i



33 / 41

Minimierung der Gesamtkosten



$$\begin{aligned} EC(q, l) &= EC(q, d_{r(1)}, d_{r(2)}, \dots, d_{r(l)}) \\ &= \sum_{i=1}^l EC(q, d_{r(i)}) \end{aligned}$$

Mimimale Gesamtkosten \rightarrow minimiere $\sum_{i=1}^l EC(q, d_{r(i)}) \rightarrow r(i)$ sollte Dokumente nach **aufsteigenden** Kosten sortieren

34 / 41

Entscheidungstheoretische Regel:

$$\begin{aligned} EC(q, d_{r(i)}) &\leq EC(q, d_{r(i+1)}) \\ \Leftrightarrow C \cdot P(R|q, d_{r(i)}) + \bar{C}(1 - P(R|q, d_{r(i)})) &\leq \\ C \cdot P(R|q, d_{r(i+1)}) + \bar{C}(1 - P(R|q, d_{r(i+1)})) & \\ \Leftrightarrow P(R|q, d_{r(i)})(C - \bar{C}) + \bar{C} &\leq \\ P(R|q, d_{r(i+1)})(C - \bar{C}) + \bar{C} & \\ \Leftrightarrow (\text{da } C < \bar{C}): P(R|q, d_{r(i)}) &\geq P(R|q, d_{r(i+1)}) \end{aligned}$$

ordne Dokumente nach der **absteigenden** Wahrscheinlichkeit der Relevanz!

35 / 41

PRP-Beispiel

System berechnet folgende Relevanzwahrscheinlichkeiten

$P(R|q, d)$:

(0.9, 0.8, 0.5, 0.4, 0.35, 0.3, 0.25, 0.2, 0.15, 0.1, 0.05, 0.0)

Benutzer schaut sich nur die ersten drei Dokumente an

1. Sei $C = 0$ und $\bar{C} = 2$.

Wie hoch sind die erwarteten Kosten für den Nutzer?

2. Erwartete Precision?

3. Erwarteter Recall?

$$\begin{aligned} 1. \quad EC(q, d) &= C \cdot P(R|q, d) + \bar{C}(1 - P(R|q, d)) = \\ &= 2 \cdot (1 - P(R|q, d)) \end{aligned}$$

$$EC(q) = 2 \cdot 0.1 + 2 \cdot 0.2 + 2 \cdot 0.5 = 1.6$$

$$2. \quad p = (0.9 + 0.8 + 0.5) / 3 = 0.73$$

$$3. \quad \sum_i P(R|q, d_i) = 4, \quad r = (0.9 + 0.8 + 0.5) / 4 = 0.55$$

36 / 41

1. vorgegebene Anzahl gefundener Dokumente
 \rightsquigarrow PRP maximiert erwarteten Recall und erwartete Precision
2. vorgegebener Recall
 \rightsquigarrow PRP maximiert erwartete Precision

37 / 41

- ▶ Minimale Kosten bei Ordnung nach fallender Relevanzwahrscheinlichkeit
- ▶ (Kosten als Optimierungskriterium für Retrieval)
- ▶ **PRP: Ordnung nach fallender Relevanzwahrscheinlichkeit liefert optimales Retrieval**
- ▶ **Dadurch theoretische Rechtfertigung für probabilistisches Retrieval**
- ▶ Für andere Modelle lässt sich dieser Zusammenhang nicht beweisen
 (z.B. bei Ranking nach fallender Ähnlichkeit beim VRM oder „optimales Relevance Feedback“ gibt es keinen direkten Zusammenhang mit Retrievalqualität)

38 / 41

BIR Exercise

Compute the values of $O(R|d_m^T)$ via eqn. 1, estimating the parameters directly as relative frequencies.

$$p_1 = \frac{8}{12} = \frac{2}{3} \quad p_2 = \frac{7}{12} \quad O(R) = \frac{12}{8} = \frac{3}{2}$$

$$s_1 = \frac{3}{8} \quad s_2 = \frac{4}{8} = \frac{1}{2}$$

$$O(R|(1, 1)) = O(R) \frac{p_1 p_2}{s_1 s_2} = \frac{28}{9}$$

$$P(R|(1, 1)) = \frac{O(R|(1, 1))}{1 + O(R|(1, 1))} = \frac{28}{37} \approx 0.76$$

$$O(R|(1, 0)) = O(R) \frac{p_1 (1 - p_2)}{s_1 (1 - s_2)} = \frac{20}{9}$$

$$P(R|(1, 0)) = \frac{O(R|(1, 0))}{1 + O(R|(1, 0))} = \frac{20}{29} \approx 0.69$$



39 / 41

Exercise for the Zhai-Lafferty Model

Given the following collection of documents:

- ▶ $d_1 = (t_1, t_1, t_1, t_2)$
- ▶ $d_2 = (t_1, t_1, t_3, t_3)$
- ▶ $d_3 = (t_1, t_2, t_2)$
- ▶ $d_4 = (t_2)$

Now consider the query $q = (t_1, t_2)$.

Compute the language model probabilities according to the Zhai-Lafferty model.

Let $\lambda = 0.5$ and assume $\alpha_d = 1$

40 / 41

Exercise for the Zhai-Lafferty Model (2)

$$P_{avg}(t_1) = \frac{6}{12} = \frac{1}{2} \quad P_{avg}(t_2) = \frac{4}{12} = \frac{1}{3}$$

$$P(t_1|d_1) = 0.5(P_{ML}(t_1|d_1) + P_{avg}(t_1)) = 0.5\left(\frac{3}{4} + \frac{1}{2}\right) = \frac{5}{8}$$

$$P(t_2|d_1) = 0.5(P_{ML}(t_2|d_1) + P_{avg}(t_2)) = 0.5\left(\frac{1}{4} + \frac{1}{3}\right) = \frac{7}{24}$$

$$P(q|d_1) = P(t_1|d_1)P(t_2|d_1) = \frac{35}{192} \approx 0.18$$

$$P(t_2|d_2) = P_{avg}(t_2) = \frac{1}{3}$$

$$P(q|d_1) = \frac{35}{192} \approx 0.18 \quad P(q|d_3) = \frac{5}{24} \approx 0.21$$

$$P(q|d_2) = \frac{1}{6} \approx 0.17 \quad P(q|d_4) = \frac{1}{3} \approx 0.33$$