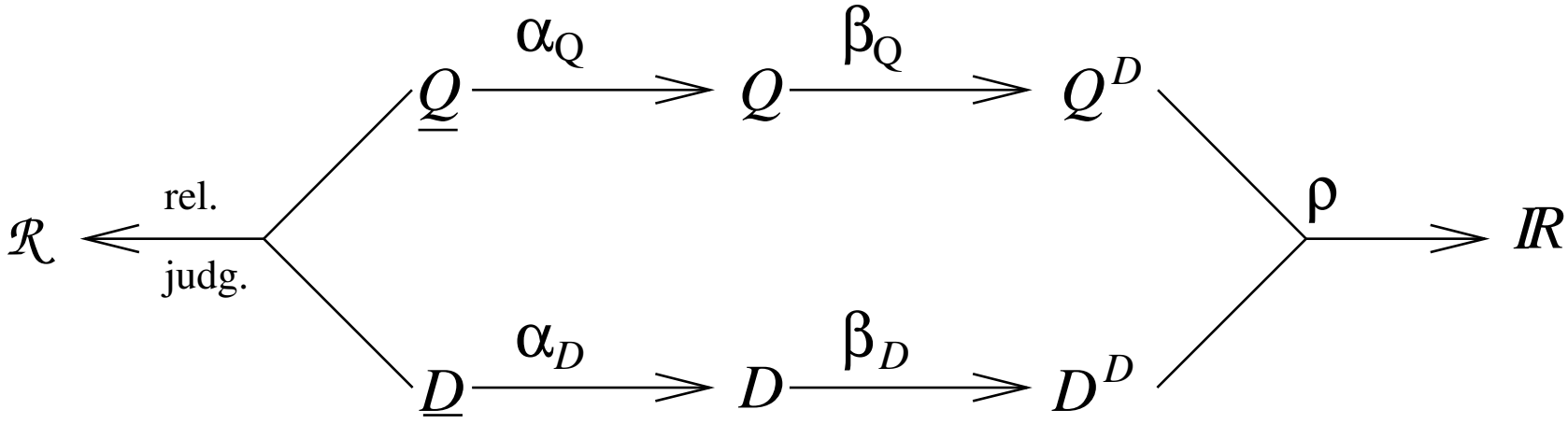
6. Probabilistische Retrievalmodelle

Norbert Fuhr

Notationen



$q \in Q$ Anfrage

$\underline{q_k} \in \underline{Q}$: Anfragerepräsentation

$q_k^D \in Q^D$: Anfragebeschreibung

$d \in D$ Dokument

$\underline{d_m} \in \underline{D}$: Dokumentrepräsentation

$d_m^D \in D^D$: Dokumentbeschreibung

$\mathcal{R}$: Relevanzskala

ϱ: Retrievalfunktion

Retrievalfunktionen für binäre Indexierung

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

$q_k \in Q$: Anfragerepräsentation — q_k^T: Menge von Fragetermen

$d_m \in D$: Dokumentrepräsentation — d_m^T: Menge von Dokumenttermen

einfache Retrievalfunktion: Coordination level match

$$\varrho_{COORD}(q_k, d_m) = |q_k^T \cap d_m^T|$$

Binary-Independence-Retrieval-Modell (BIR):
weise Fragetermen Gewichte zu

$$\varrho_{BIR}(q_k, d_m) = \sum_{t_i \in q_k^T \cap d_m^T} c_{ik}$$

Probabilistische Grundlagen des BIR-Modells

Grundlegende mathematische Techniken zur Herleitung der probabilistischen Retrievalmodelle:

1. Anwendung des Bayes'schen Theorems:

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

2. Benutzung von Chancen statt Wahrscheinlichkeiten, wobei

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

Herleitung des BIR-Modells

Abschätzung von $O(R|q_k, d_m^T)$
= Chance, dass ein Dokument mit einer Menge von Termen d_m^T relevant zur Anfrage q_k ist

Repräsentation des Dokuments d_m als Vektor mit binären Komponenten $\vec{x} = (x_1, \ldots, x_n)$ wobei

$$x_i = \begin{cases} 1, & \text{falls } t_i \in d_m^T \\ 0, & \text{sonst} \end{cases}$$

Anwenden des Bayes'schen Theorems:

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

$P(R|q_k)$: Wahrscheinlichkeit, dass ein arbiträres Dokument relevant ist zu q_k
$P(\vec{x}_m|R, q_k)$: Wahrscheinlichkeit, dass ein arbiträres, relevantes Dokument den Termvektor $\vec{x}$ besitzt
$P(\vec{x}_m|\bar{R}, q_k)$: Wahrscheinlichkeit, dass ein arbiträres, nicht-relevantes Dokument den Termvektor $\vec{x}$ besitzt

Annahme der "Linked dependence":

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

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

Aufteilen nach Vorkommen/Fehlen von Termen im aktuellen Dokument:

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

$p_{ik} = P(x_i{=}1|R,q_k)$: Wahrscheinlichkeit, dass t_i in einem arbiträren relevanten Dokument vorkommt.
$q_{ik} = P(x_i{=}1|\bar{R},q_k)$: Wahrscheinlichkeit, dass t_i in einem arbiträren nicht-relevanten Dokument vorkommt

Annahme, dass $p_{ik} = q_{ik}$ für alle $t_i \notin q_k^T$

$$
\begin{aligned}
O(R|q_k, d_m^T) &= O(R|q_k) \prod_{t_i \epsilon d_m^T \cap q_k^T} \frac{p_{ik}}{q_{ik}} \cdot \prod_{t_i \epsilon q_k^T \setminus d_m^T} \frac{1-p_{ik}}{1-q_{ik}} \\
&= O(R|q_k) \prod_{t_i \epsilon d_m^T \cap q_k^T} \frac{p_{ik}}{q_{ik}} \cdot \prod_{t_i \epsilon d_m^T \cap q_k^T} \frac{1-q_{ik}}{1-p_{ik}} \\
&\quad \cdot \prod_{t_i \epsilon d_m^T \cap q_k^T} \frac{1-p_{ik}}{1-q_{ik}} \cdot \prod_{t_i \epsilon q_k^T \setminus d_m^T} \frac{1-p_{ik}}{1-q_{ik}} \\
&= O(R|q_k) \prod_{t_i \epsilon d_m^T \cap q_k^T} \frac{p_{ik}(1-q_{ik})}{q_{ik}(1-p_{ik})} \cdot \prod_{t_i \epsilon q_k^T} \frac{1-p_{ik}}{1-q_{ik}}
\end{aligned}
$$

Nur das erste Produkt ist bezüglich einer gegebenen Anfrage q_k für unterschiedliche Dokumente *nicht* konstant $\longrightarrow$
Betrachte daher nur dieses Produkt für das Ranking

Übergang zum Logarithmus (ordnungserhaltend):

$$c_{ik} = \log \frac{p_{ik}(1 - q_{ik})}{q_{ik}(1 - p_{ik})}$$

Retrievalfunktion:

$$\varrho_{BIR}(q_k, d_m) = \sum_{t_i \epsilon d_m^T \cap q_k^T} c_{ik}$$

Anwendung des BIR-Modells

Parameterabschätzung für q_{ik}

$q_{ik} = P(x_i{=}1|\bar{R}, q_k)$:
(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
$q_{ik} = \frac{n_i}{N}$

Parameterabschätzung für p_{ik}

$p_{ik} = P(x_i{=}1|R, q_k)$:
(Wahrscheinlichkeit, dass t_i in einem arbiträren relevanten Dokument vorkommt)

1. benutze globalen Wert p für alle p_{ik}s

$\longrightarrow$ Termgewichtung nach inverser Dokumentenhäufigkeit (IDF)

$$\begin{aligned} c_{ik} &= \log \frac{p}{1-p} + \log \frac{1-q_{ik}}{q_{ik}} \\ &= c_p + \log \frac{N-n_i}{n_i} \end{aligned}$$

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

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

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_k

r_i: # relevante Dokumente mit dem Term t_i

$p_{ik} = P(t_i|R, q_k) \approx \frac{r_i}{r}$

verbesserte Abschätzungen (mehr in späterem Abschnitt):

$$p_{ik} \approx \frac{r_i + 0.5}{r + 1}$$

Beispiel für BIR

d_m	$r(d_m)$	x_1	x_2	$P(R\|\vec{x})$	BIR
d_1	R	1	1		
d_2	R	1	1		
d_3	R	1	1	0.80	0.76
d_4	R	1	1		
d_5	N	1	1		
d_6	R	1	0		
d_7	R	1	0		
d_8	R	1	0	0.67	0.69
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		
d_{13}	R	0	1		
d_{14}	R	0	1	0.50	0.48
d_{15}	N	0	1		
d_{16}	N	0	1		
d_{17}	N	0	1		
d_{18}	R	0	0		
d_{19}	N	0	0	0.33	0.40
d_{20}	N	0	0		

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

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

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

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

Entscheidungstheoretische Regel:

$$EC(q, d_{r(i)}) \leq EC(q, d_{r(i+1)}) \iff$$

$$\begin{aligned} C \cdot P(R|q, d_{r(i)}) + \bar{C}(1 - P(R|q, d_{r(i)})) \quad &\leq \\ C \cdot P(R|q, d_{r(i+1)}) + \bar{C}\left(1 - P(R|q, d_{r(i+1)})\right) & \end{aligned}$$

$$\iff \text{(da } C < \bar{C}\text{):} \qquad P(R|q, d_{r(i)}) \geq P(R|q, d_{r(i+1)}).$$

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

Rechtfertigung über Effektivitätsmaße

für je zwei Ereignisse a, b, liefert das Bayes'sche Theorem die folgenden monotonen Transformationen von $P(a|b)$: (siehe Herleitung des BIR-Modells)

$$
\begin{aligned}
O(a|b) &= \frac{P(b|a)P(a)}{P(b|\bar{a})P(\bar{a})} \\
\log O(a|b) &= \log \frac{P(b|a)}{P(b|\bar{a})} + \log O(a) \\
\operatorname{logit} P(a|b) &= \log \frac{P(b|a)}{P(b|\bar{a})} + \operatorname{logit} P(a)
\end{aligned}
$$

mit $\operatorname{logit} P(x) = \log O(x)$

$$\rho = P(\text{gef. Dokument}|\text{rel. Dokument})$$
$$\phi = P(\text{gef. Dokument}|\text{nichtrel. Dokument})$$
$$\pi = P(\text{rel. Dokument}|\text{gef. Dokument})$$
$$\gamma = P(\text{rel. Dokument})$$
$$\rho(d_i) = P(\text{Dokument ist } d_i|\text{rel. Dokument})$$
$$\phi(d_i) = P(\text{Dokument ist } d_i|\text{nichtrel. Dokument})$$
$$\pi(d_i) = P(\text{rel. Dokument}|\text{Dokument ist } d_i)$$

(Wahrscheinlichkeit der Relevanz)

S Menge der gefundenen Dokumente

$$
\begin{aligned}
\rho &= \sum_{d_i \in S} \rho(d_i) \\
\phi &= \sum_{d_i \in S} \phi(d_i)
\end{aligned}
$$

$$
\begin{aligned}
\text{logit } \pi(d_i) &= \log \frac{\rho(d_i)}{\phi(d_i)} + \text{logit } \gamma \\
\rho(d_i) &= x_i \cdot \phi(d_i) \qquad \text{mit} \\
x_i &= \exp(\text{logit } \pi(d_i) - \text{logit } \gamma)
\end{aligned}
$$

1. Abbruch vorgegeben durch ϕ (Fallout)

$$\begin{aligned} \phi &= \sum_{d_i \in S} \phi(d_i) \\ \rho &= \sum_{d_i \in S} \rho(d_i) = \sum_{d_i \in S} \phi(d_i) \cdot \exp(\text{logit } \pi(d_i) - \text{logit } \gamma) \end{aligned}$$

   $\rightsquigarrow$ maximiere ρ (Recall) durch Hinzunahme der Dokumente mit den höchsten Werten für $\pi(d_i)$
   $\hat{=}$ ordne nach Wahrscheinlichkeit der Relevanz

2. Abbruch durch # Dokumente gefunden
   $\rightsquigarrow$ maximiere erwarteten Recall, minimiere erwarteten Fallout

3. Abbruch vorgegeben durch ρ (Recall)
   $\rightsquigarrow$ minimiere Fallout

$$\text{logit } \pi = \log(\rho/\phi) + \text{logit } \gamma$$

4. erwartete Precision wird für gegebenen Recall / Fallout / # gefundener Dokumente maximiert

PRP für mehrwertige Relevanzskalen

n Relevanzwerte $R_1 < R_2 < \ldots < R_n$
entsprechende Kosten für das Retrieval eines Dokuments:
$C_1, C_2, \ldots, C_n$.
ordne Dokumente nach ihren erwarteten Kosten

$$EC(q, d_m) = \sum_{l=1}^{n} C_l \cdot P(R_l|q, d_m).$$

Vergleich mit dem binären Fall:

- nicht-binäre Skala entspricht eher dem Benutzerempfinden
- $n - 1$ Schätzungen $P(R_l|q, d_m)$ werden benötigt
- Kostenfaktoren C_l müssen bekannt sein
- widerspricht bisher experimentellen Ergebnissen

Kombination von probabilistischen und Fuzzy-Retrieval

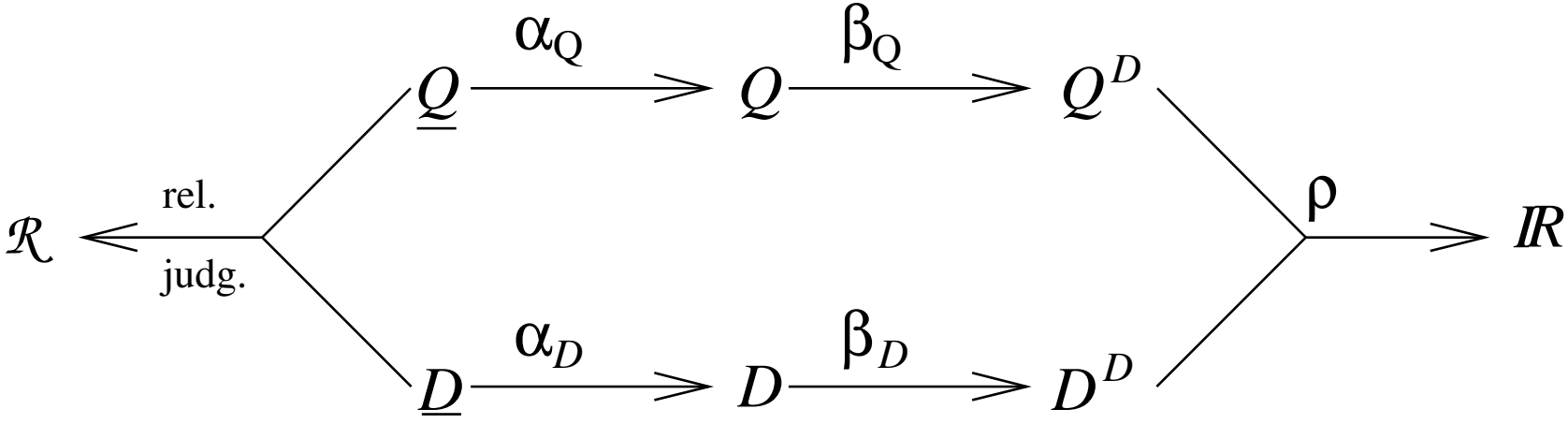
Fuzzy-Retrieval:

- benutzt *Grad der Relevanz* statt binärer Skala
- System versucht Grad der Relevanz für ein Anfrage-Dokument-Paar zu berechnen

Kombination:

- kontinuierliche Relevanzskala: $r \epsilon [0, 1]$
- ersetze Wahrscheinlichkeitsverteilung $P(R_l|q, d_m)$ durch Dichtefunktion $p(r|q, d_m)$
- ersetze Kostenfaktoren C_l durch Kostenfunktion $c(r)$.

Konzeptuelles Modell



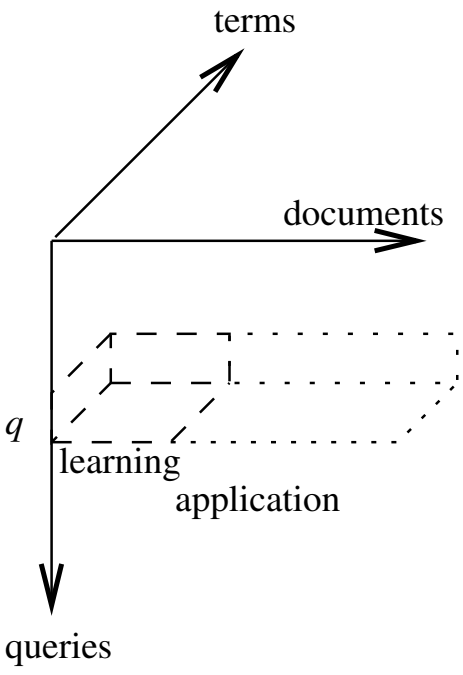
Repräsentationen und Beschreibungen im BIR-Modell

- Anfragerepräsentationen $q_k = (q_k^T, q_k^J)$:
  Menge von Anfragetermen q_k^T +
  Menge von Rlevanzurteilen $q_k^J = \{(\underline{d_m}, r(\underline{d_m}, \underline{q_k}))\}$
- Anfragebeschreibungen $q_k^D = \{(t_i, c_{ik})\}$:
  Menge der Anfrageterme mit zugehörigen Gewichten
- Dokumentenrepräsentation $d_m = d_m^T$
  Menge der Terme
- Dokumentenbeschreibung d_m^D = Dokumentenrepräsentation d_m^T

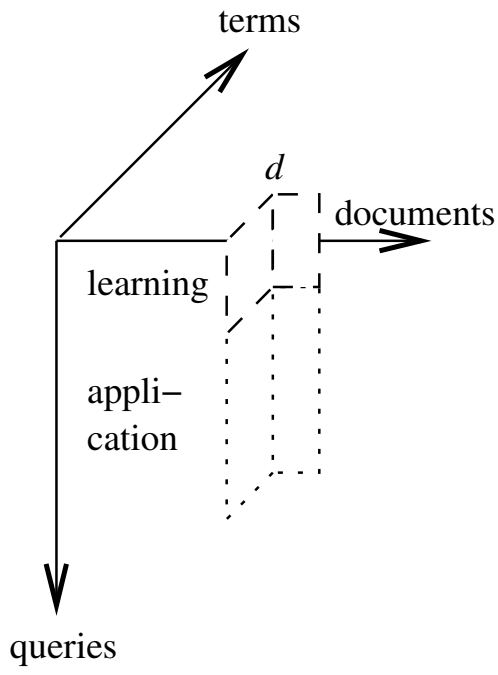
Entwicklungsrichtungen für probabilistische IR-Modelle:

1. Optimierung der Retrievalqualität für feste Repräsentationen
   (z.B. durch andere Abhängigkeitsannahmen als im BIR-Modell)
2. Modelle für detaillierte Repräsentationen
   (z.B. Dokumente als Multimengen von Termen, Phrasen
   zusätzlich zu Worten)

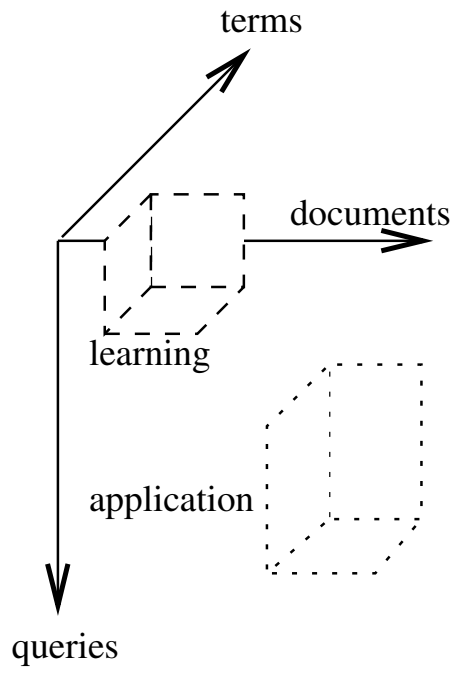
Parameterlernen im IR



query–related
learning



document–related
learning



description–related
learning

Lernansätze im IR

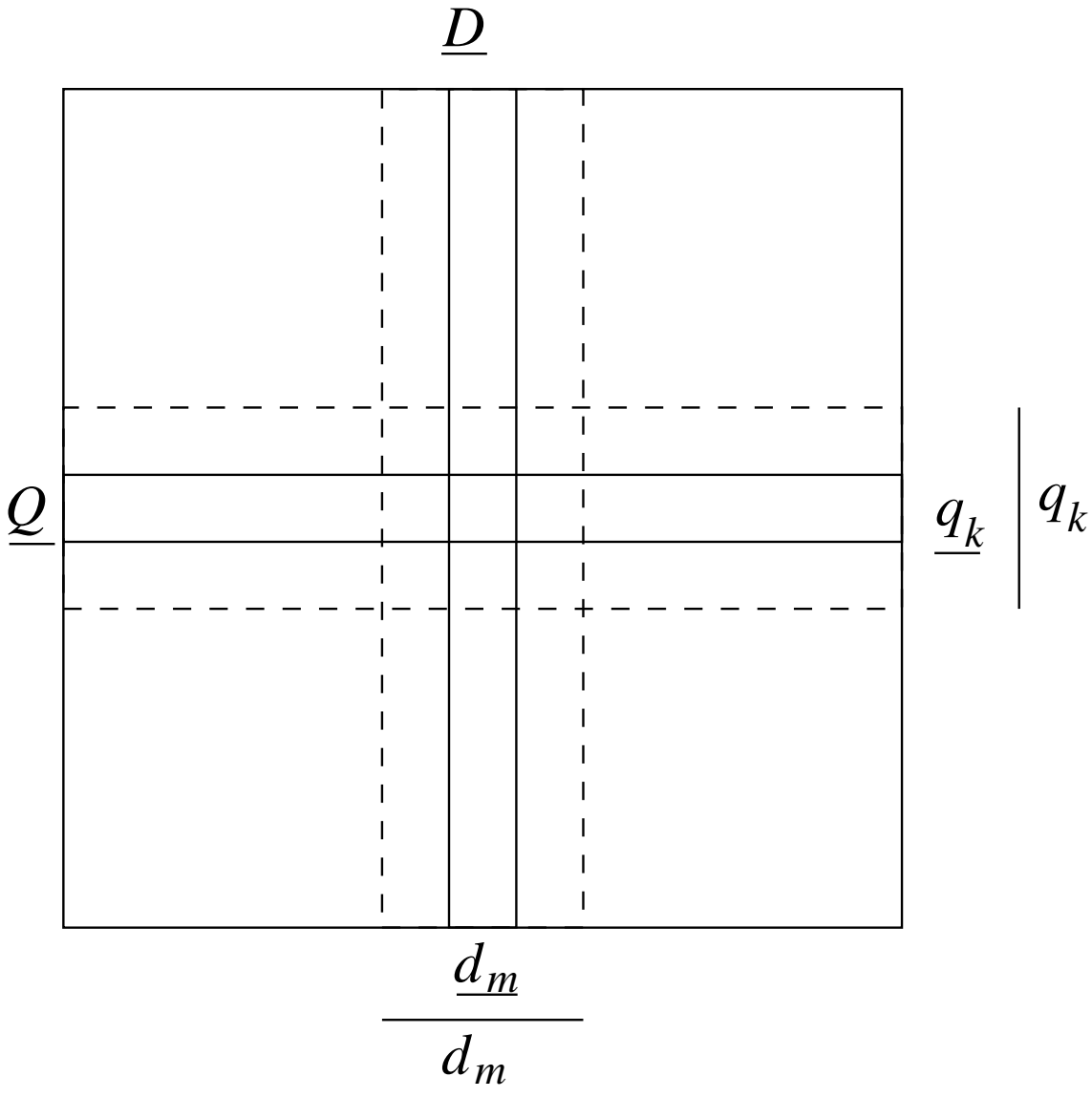
Ereignisraum

Ereignisraum: $\underline{Q} \times \underline{D}$
einzelnes Element: Frage-Dokument-Paar $(\underline{q_k}, \underline{d_m})$
alle Elemente sind gleichwahrscheinlich

Relevanzurteile $(\underline{q_k}, \underline{d_m}) \epsilon \mathcal{R}$
Relevanzurteile für verschiedene Dokumente bzgl. der gleichen Anfrage sind unabhängig voneinander

Wahrscheinlichkeit der Relevanz $P(R|q_k, d_m)$:
Wahrscheinlichkeit, dass ein Element (q_k, d_m) relevant ist

- betrachte Kollektionen als Ausschnitt von möglicherweise unendlichen Mengen
- schlechte Repräsentation von gefundenen Objekten:
  eine einzelne Repräsentation kann für mehrere verschiedene Objekte stehen



Ereignisraum der Relevanzmodelle