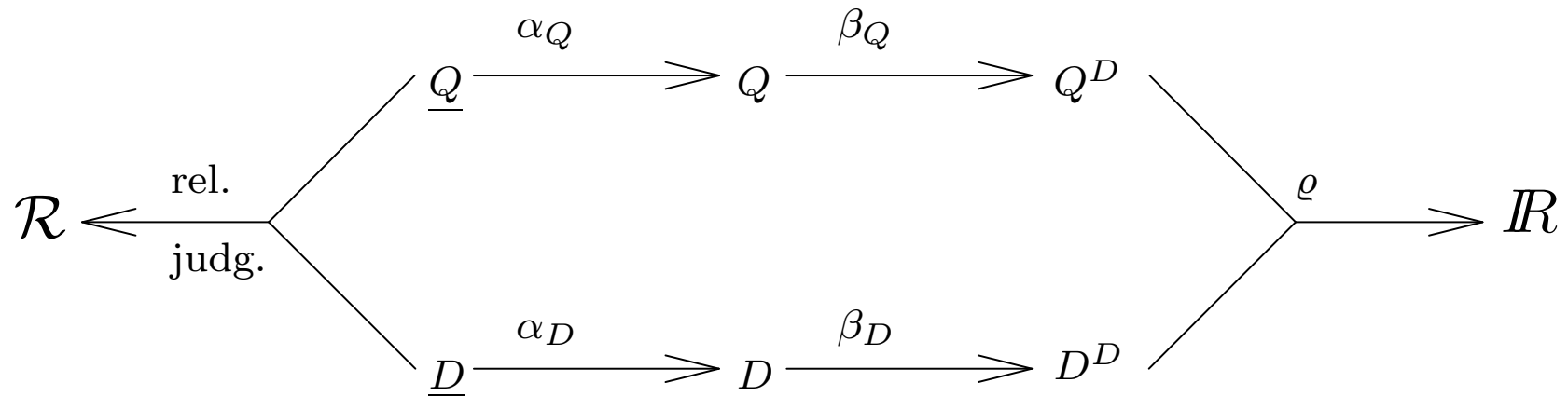


6 Probabilistic Retrieval Models

- Notations
- Binary Independence Retrieval model
- Probability Ranking Principle

6.1 Notations



$\underline{q} \in \underline{Q}$ query

$q_k \in Q$: query representation

$q_k^D \in Q^D$: query description

\mathcal{R} : relevance scale

ϱ : retrieval function

$\underline{d} \in \underline{D}$ document

$d_m \in D$: document representation

$d_m^D \in D^D$: document description

6.2 Binary independence retrieval model

6.2.1 Retrieval functions for binary indexing

represent queries and documents as sets of terms

$T = \{t_1, \dots, t_n\}$ set of terms in the collection

$q_k \in Q$: query representation q_k^T : set of query terms

$d_m \in D$: document representation d_m^T : set of document terms

simple retrieval function: coordination level match

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

Binary independence retrieval (BIR) model:

assign weights to query terms

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

6.2.2 Probabilistic foundation of the BIR model

Basic techniques for the derivation of probabilistic models:

1. application of Bayes' theorem:

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

,

2. usage of odds instead of probabilities, where

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

Derivation of the BIR model

Estimation of $O(R|q_k, d_m^T)$

= odds that document with set of terms d_m^T will be relevant to q_k

represent document d_m as binary vector $\vec{x} = (x_1, \dots, x_n)$ with

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

Apply Bayes' Theorem:

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

$P(R|q_k)$: prob. that arbitrary doc. will be relevant to q_k (generality of q_k)

$P(\vec{x}_m|R, q_k)$: prob. that arbitrary relevant doc. will have term vector \vec{x}

$P(\vec{x}_m|\bar{R}, q_k)$: prob. that arbitrary nonrelevant doc. will have term vector \vec{x}

Linked dependence assumption:

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

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

split according to presence/absence of terms in the current document:

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

$p_{ik} = P(x_i=1|R, q_k)$: prob. that t_i occurs in arbitrary relevant doc.

$q_{ik} = P(x_i=1|\bar{R}, q_k)$ prob. that t_i occurs in arbitrary nonrelevant doc.

assume that $p_{ik} = q_{ik}$ for all $t_i \notin q_k^T$

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

Only first product varies for different documents with respect to the same request

$q_k \longrightarrow$

regard only this product for ranking

use logarithm:

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

retrieval function:

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

6.2.3 Application of the BIR model

Parameter estimation for q_{ik}

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

(probability that t_i occurs in arbitrary nonrelevant document)

assume that number of nonrelevant documents \approx collection size

N – collection size

n_i – # documents with term t_i

$$q_{ik} = \frac{n_i}{N}$$

Parameter estimation for p_{ik}

$$p_{ik} = P(x_i=1|R, q_k):$$

(probability that t_i occurs in arbitrary relevant document)

1. assume global value p for all p_{ik} s

—→ term weighting by inverse document frequency (IDF)

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

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

often used: $p = 0.5 \rightarrow c_p = 0$

2. relevance feedback:

initial ranking with IDF formula

present top ranking documents to the user

(about 10...20)

user gives binary relevance judgements: relevant/non-relevant

r : # documents judged relevant for request q_k

r_i : # relevant documents with term t_i

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

improved estimates (see parameter estimation methods):

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

BIR example

d_m	$r(d_m)$	x_1	x_2	$P(R \vec{x})$	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_{12}	R	0	1	0.50	0.48
d_2	R	1	1			d_{13}	R	0	1		
d_3	R	1	1			d_{14}	R	0	1		
d_4	R	1	1			d_{15}	N	0	1		
d_5	N	1	1			d_{16}	N	0	1		
d_6	R	1	0	0.67	0.69	d_{17}	N	0	1	0.33	0.40
d_7	R	1	0			d_{18}	R	0	0		
d_8	R	1	0			d_{19}	N	0	0		
d_9	R	1	0			d_{20}	N	0	0		
d_{10}	N	1	0								
d_{11}	N	1	0								

6.3 The Probability Ranking Principle (PRP)

perfect retrieval:

rank all relevant documents ahead of any nonrelevant one
relates to objects itself, only possible with complete relevance information

optimum retrieval: relates to representations (as any IR system does)

Probability Ranking Principle (PRP)

defines optimum retrieval for probabilistic models:

rank documents according to decreasing probability of relevance

6.3.1 Decision-theoretic justification of the PRP

\bar{C} : costs for the retrieval of a nonrelevant document

C : costs for the retrieval of a relevant document.

expected costs for the retrieval of a document d_j :

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

decision-theoretic rule:

retrieve document d_m with minimum expected costs, i.e. if

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

for any other document d_j in the collection (not yet retrieved)

$$\iff (\text{since } C < \bar{C}): \quad P(R|q, d_m) \geq P(R|q, d_j).$$

rank documents according to their decreasing probability of relevance!

6.3.2 PRP for multivalued relevance scales

n relevance values $R_1 < R_2 < \dots < R_n$

corresponding costs for the retrieval of a document: C_1, C_2, \dots, C_n .

rank documents according to their expected costs

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

comparison with binary case:

- nonbinary scale more appropriate for user
- $n - 1$ estimates $P(R_l|q, d_m)$ required
- cost factors C_l must be known
- contradicting experimental evidence so far

Combination of probabilistic and fuzzy retrieval

Fuzzy retrieval:

- uses *degree of relevance* instead of binary scale
- system aims at computing a degree of relevance for a query-document pair

Combination:

- continuous relevance scale: $r \in [0, 1]$
- replace probability distribution $P(R_l|q, d_m)$ by density function $p(r|q, d_m)$
- replace cost factors C_l by cost function $c(r)$.

6.4 Summary: Probabilistic retrieval

- + based on solid theoretical model
(all assumptions made explicit)
- + model refers to retrieval quality
- + gives high retrieval quality
- more sophisticated probabilistic models need training data for new collections