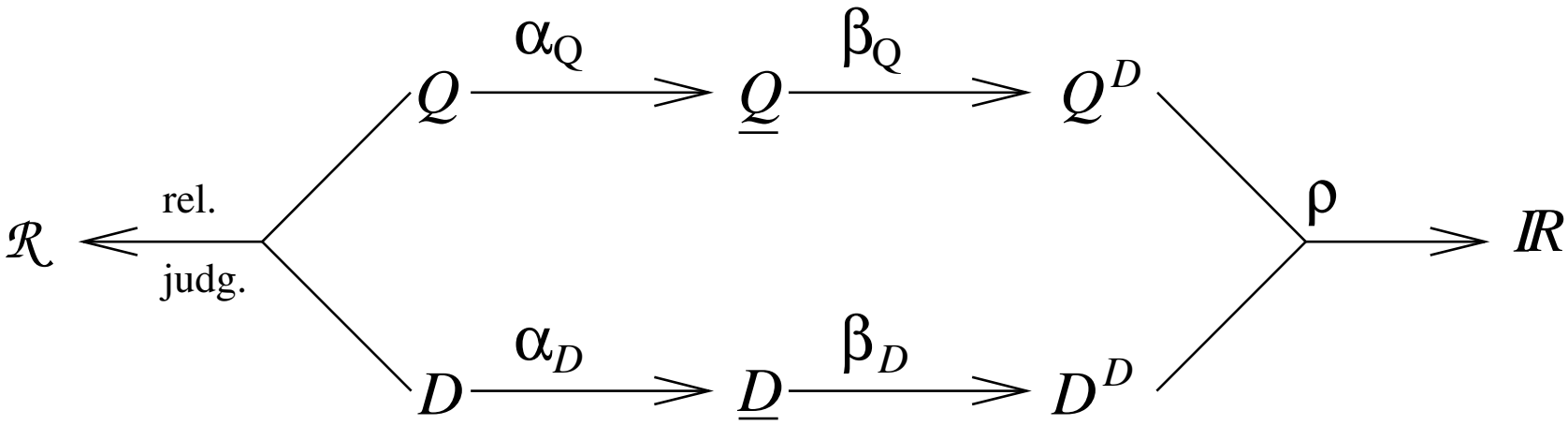
6. Probabilistic Retrieval Models

Norbert Fuhr

Notations



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

$q_k \in Q$: query representation

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

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

$d_m \in D$: document representation

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

$\mathcal{R}$: relevance scale

ϱ: retrieval function

Retrieval functions for binary indexing

represent queries and documents as sets of terms
$T = \{t_1, \ldots, t_n\}$ set of terms in the collection

$q_k \in Q$: query representation

$d_m \in D$: document representation

q_k^T: set of query terms

d_m^T: set of document terms

simple retrieval function: Coordination level match

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

Binary independence retrieval (BIR) model:
assign weights to query terms

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

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 $\vec{x} = (x_1, \ldots, x_n)$ 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
$P(\vec{x}_m|R, q_k)$: 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 $t_i \notin q_k^T$

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

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:

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

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
$\longrightarrow$ term weighting by inverse document frequency (IDF)

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

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

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

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_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		

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

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:

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

Total cost for retrieval:
(assuming that user looks at first l documents, where l is not known in advance)
$r(i)$: ranking function, determines index of document to be placed at rank i

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

Mimimum total costs $\rightarrow$ minimize $\sum_{i=1}^{l} EC(q, d_{r(i)}) \rightarrow$ $r(i)$ should order documents by ascending costs

Decision-theoretic rule:

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

$$C \cdot P(R|q, d_{r(i)}) + \bar{C}(1 - P(R|q, d_{r(i)})) \leq C \cdot P(R|q, d_{r(i+1)}) + \bar{C}\left(1 - P(R|q,\right.$$

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

rank documents according to their decreasing probability of relevance!

Justification based on effectiveness measures

for any two events a, b, Bayes' theorem yields
the following monotonic transformations of $P(a|b)$:
(see derivation of BIR model)

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

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

$$
\begin{aligned}
\rho &= P(\text{doc.retrieved}|\text{doc.relevant}) \\
\phi &= P(\text{doc.retrieved}|\text{doc.nonrelevant}) \\
\pi &= P(\text{doc.relevant}|\text{doc.retrieved}) \\
\gamma &= P(\text{doc.relevant}) \\
\rho(d_i) &= P(\text{doc.is } d_i|\text{doc.relevant}) \\
\phi(d_i) &= P(\text{doc.is } d_i|\text{doc.nonrelevant}) \\
\pi(d_i) &= P(\text{doc.relevant}|\text{doc.is } d_i) \text{ (probability of relevance)} \\
S &\text{ set of retrieved documents}
\end{aligned}
$$

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

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

1. cutoff defined by ϕ (fallout)

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

   $\rightsquigarrow$ maximize ρ (recall) by including docs with highest values of $\pi(d_i)$
   $\hat{=}$ rank according to prob. of relevance

2. given # documents retrieved
   $\rightsquigarrow$ maximize expected recall, minimize expected fallout

3. cutoff defined by ρ (recall)
   $\rightsquigarrow$ minimize fallout

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

4. expected precision is maximized for given recall / fallout / # documents retrieved

PRP for multivalued relevance scales

n relevance values $R_1 < R_2 < \ldots < R_n$
corresponding costs for the retrieval of a document: $C_1, C_2, \ldots, C_n$.
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

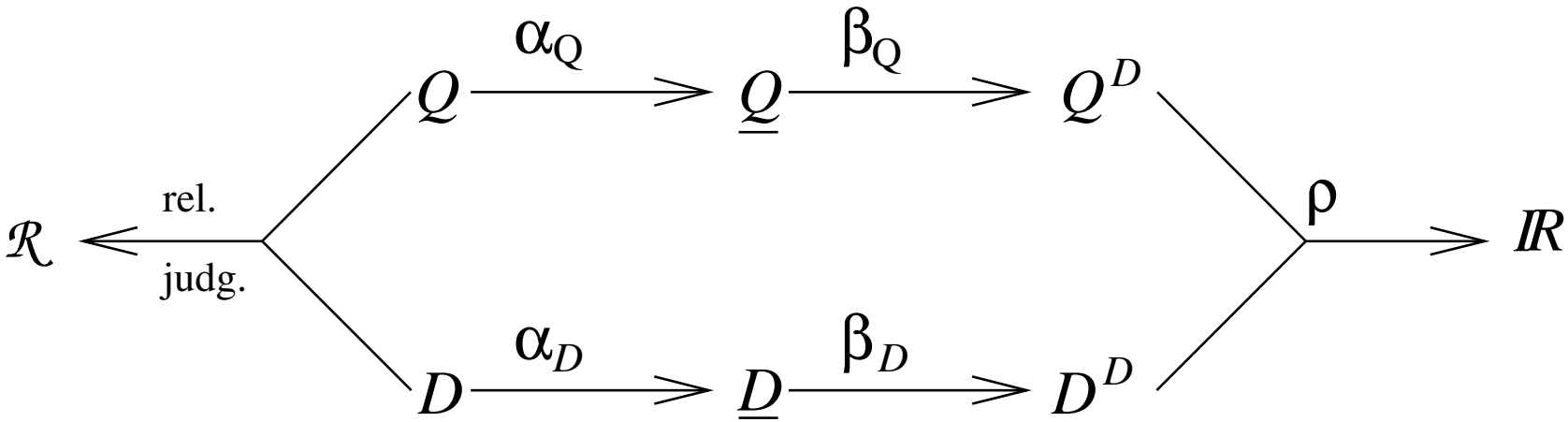
Fuzzy retrieval:

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

Combination:

- continuous relevance scale: $r \epsilon [0, 1]$
- 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 $c(r)$.

Conceptual model



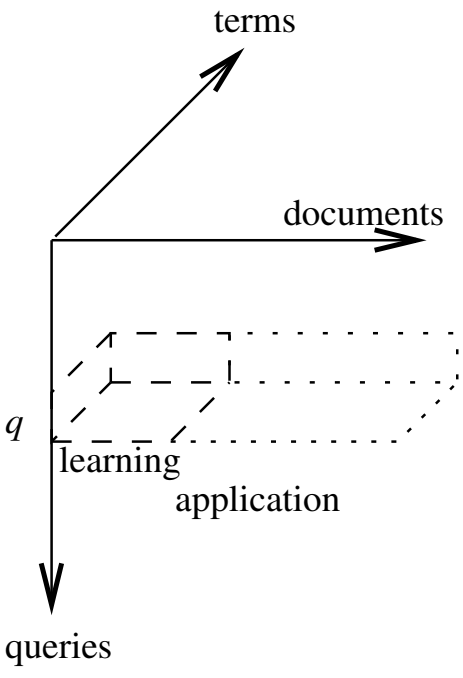
Representations and descriptions of the BIR model

- query representation $q_k = (q_k^T, q_k^J)$:
  set of query terms q_k^T +
  set of relevance judgements $q_k^J = \{(\underline{d_m}, r(\underline{d_m}, \underline{q_k}))\}$
- query description $q_k^D = \{(t_i, c_{ik})\}$:
  set of query terms with associated weights
- document representation $d_m = d_m^T$
  set of terms
- document description $d_m^D =$ document representation d_m^T

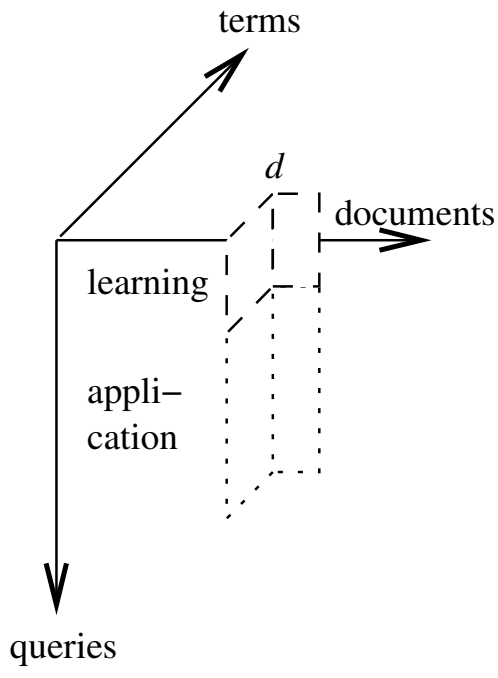
Directions in the development of probabilistic IR models:

1. Optimization of retrieval quality for a fixed representation
   (e.g. different dependence assumptions than in the BIR model)
2. models for more detailed representations
   (e.g. documents as bags of terms, phrases in addition to words)

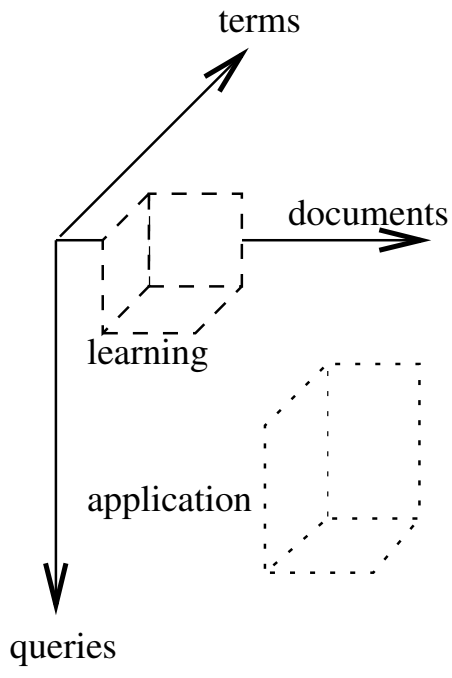
Parameter learning in IR



query–related
learning



document–related
learning



description–related
learning

Learning approaches in IR

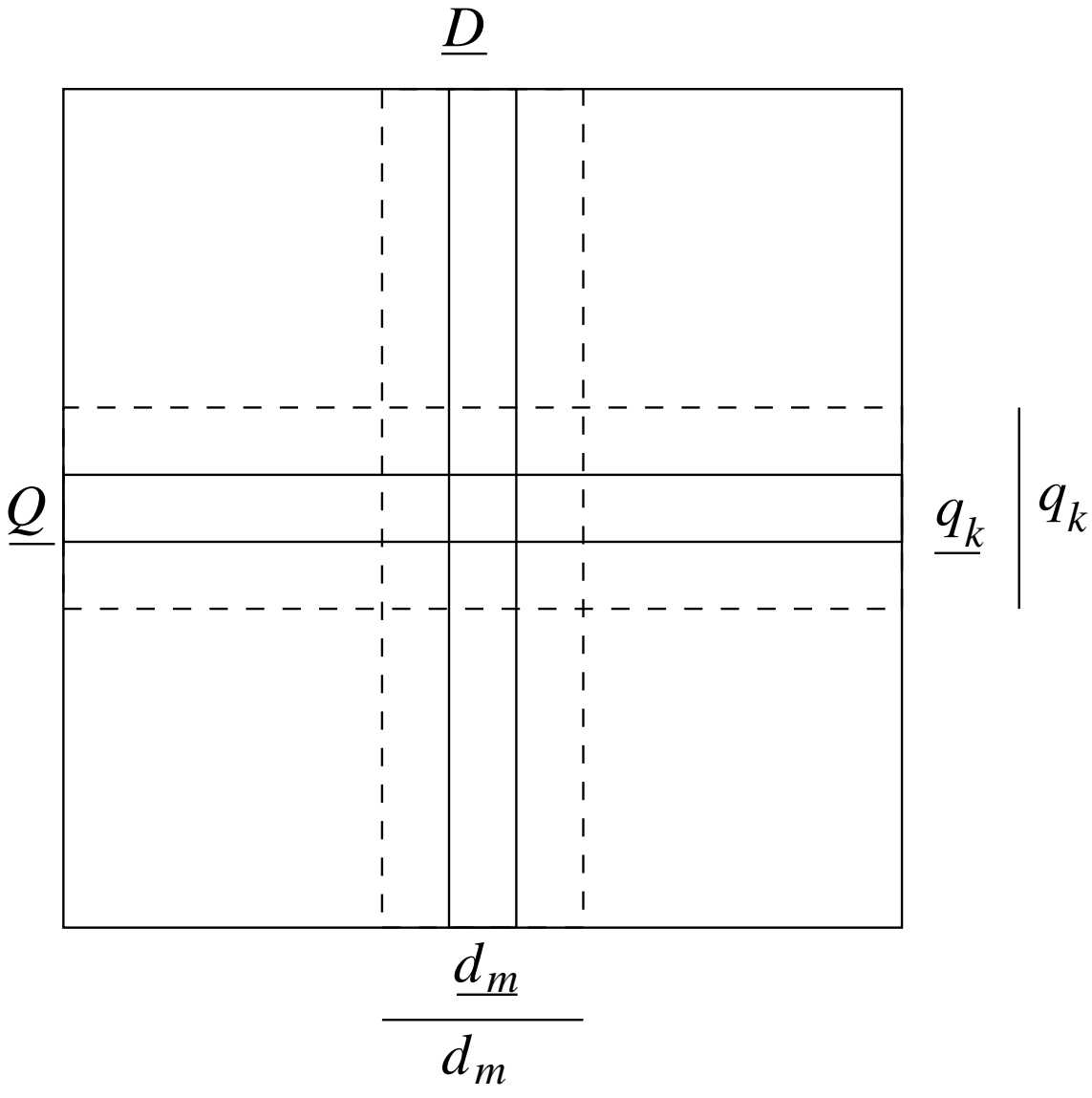
Event space

Event space: $\underline{Q} \times \underline{D}$
single element: query-document pair $(\underline{q_k}, \underline{d_m})$
all elements are equiprobable

relevance judgement $(\underline{q_k}, \underline{d_m}) \epsilon \mathcal{R}$
relevance judgements for different documents w.r.t. the same query are independent of each other

Probability of relevance $P(R|q_k, d_m)$:
probability of a an element of (q_k, d_m) being relevant

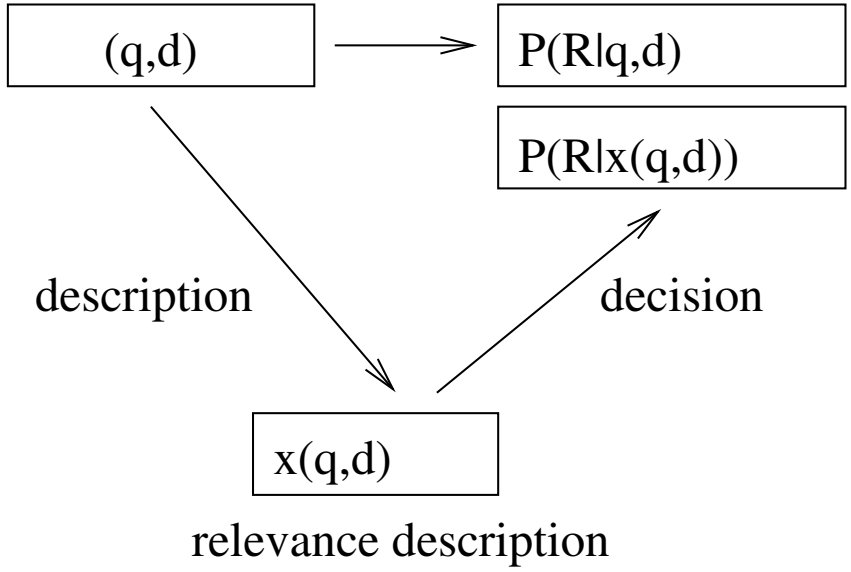
- regard collections as samples of possibly infinite sets
- poor representation of retrieval objects:
  single representation may stand for a number of different objects.



Event space of relevance models

Optimum polynomial retrieval functions
Basic concepts

regard retrieval as (probabilistic) classification task
(classify objects into one of n classes)
objects: query-document pairs $(\underline{q}_k, \underline{d}_m)$
classes: relevance values $R_l \in \mathcal{R} = \{R_1, \ldots, R_n\}$
description-oriented approach:
learning strategy abstracting from specific queries, documents and terms



1. **description step**
   map query-document pairs onto a feature vector $\vec{x} = \vec{x}(q_k, d_m)$
2. **decision step**
   apply classification functions $e_l(\vec{x})$ for estimating
   $P(R_l|\vec{x}(q_k, d_m)), l = 1, \ldots, n$
   classification functions are derived from learning sample with relevance judgements

Description step

Example:

element	description
x_1	# descriptors common to query and document
x_2	log(# descriptors common to query and document)
x_3	highest indexing weight of a common descriptor
x_4	lowest indexing weight of a common descriptor
x_5	# common descriptors with weight ≥ 0.15
x_6	# non-common descriptors with weight ≥ 0.15
x_7	# descriptors in the document with weight ≥ 0.15
x_8	log$\sum$(indexing weights of common descriptors)
x_9	log(# descriptors in the query)
x_{10}	log(min(size of output set, 100))
x_{11}	$= 1$, if size of output set > 100
x_{12}	$= 1$, if request about nuclear physics

Decision step

represent relevance judgement $R_l = r(q_k, d_m)$ by a vector $\vec{y} = (y_1, \ldots, y_n)$ with

$$y_i = \begin{cases} 1, & \text{if } i = l \\ 0 & \text{otherwise.} \end{cases}$$

seek for regression function $\vec{e}_{opt}(\vec{x})$
which yields optimum approximations $\hat{\vec{y}}$ of the class vectors $\vec{y}$.
optimizing criterion: minimum squared errors

$$E(|\vec{y} - \vec{e}_{opt}(\vec{x})|^2) \stackrel{!}{=} \min .$$

$\longrightarrow$ $\vec{e}_{opt}(\vec{x})$ yields probabilistic estimates $P(R_l|\vec{x})$ in the components of $\hat{\vec{y}}$
variation problem cannot be solved in its general form
$\longrightarrow$ restrict search to predefined class of functions
general variation problem $\longrightarrow$ parameter optimization task.

resulting functions yield least squares approximations of $\vec{e}_{opt}$: approximation with respect to the expression

$$E(|\vec{y} - \hat{\vec{y}}\,|^2) \stackrel{!}{=} \min$$

yields the same result as an optimization fulfilling the condition

$$E(|E(\vec{y}\,|\vec{x}\,) - \hat{\vec{y}}\,|^2) \stackrel{!}{=} \min\,.$$

$\longrightarrow$ parameter optimiziation yields least squares approximations of $P(R_l|\vec{x}(q_k, d_m))$.

Least square polynomials (LSP) approach:

polynomials with a predefined structure as function classes
define polynomial structure

$$\vec{v}\,(\vec{x}) = (v_1, \ldots, v_L)$$

with

$$\vec{v}\,(\vec{x}) = (1, x_1, x_2, \ldots, x_N, x_1^2, x_1 x_2, \ldots)$$

$N =$ number of dimensions of $\vec{x}$

class of polynomials is given by the components

$$x_i^l \cdot x_j^m \cdot x_k^n \cdot \ldots (i, j, k, \ldots \in [1, N]; l, m, n, \ldots \geq 0)$$

(mostly linear and quadratic polynomials regarded)

regression function:

$$\vec{e}\,(\vec{x}) = A^T \cdot \vec{v}\,(\vec{x})$$

where $A = (a_{il})$ with i=1, . . . , L; l=1, . . . , n is the coefficient matrix $P(R_l|\vec{x})$ is approximated by the polynomial

$$\begin{aligned} e_l(\vec{x}) \quad = \quad & a_{1l} + a_{2l} \cdot x_1 + a_{3l} \cdot x_2 + \ldots + a_{N+1,l} \cdot x_N + \\ & + a_{N+2,l} \cdot x_1^2 + a_{N+3,l} \cdot x_1 \cdot x_2 + \ldots \end{aligned}$$

coefficient matrix is computed by solving the equation system

$$E(\vec{v} \cdot \vec{v}^{\,T}) \cdot A = E(\vec{v} \cdot \vec{y}^{\,T}). \qquad (1)$$

Development of an LSP function:
1. Statistical evaluation of a learning sample

- use representative sample of request-document relationships together with relevance judgements
- derive pairs $(\vec{x}, \vec{y}\,)$
- compute empirical momental matrix M:

$$M = (\overline{\vec{v} \cdot \vec{v}^{\,T}} \ \overline{\vec{v} \cdot \vec{y}^{\,T}}).$$

(M contains both sides of the equation system)

2. Computation of the coefficient matrix

by means of the Gauss-Jordan algorithm
choose coefficient which maximizes the reduction of the overall error s^2
matrix M before the ith elimination step:
$M^{(i)} = (m_{lj}^{(i)})$ with l=1, . . . , n; j=1, . . . , $L + n$
reduction $d_j^{(i)}$ achieved by choosing component j:

$$d_j^{(i)} = \frac{1}{m_{jj}^{(i)^2}} \cdot \sum_{l=L+1}^{L+n} m_{jl}^{(i)^2}. \qquad (2)$$

procedure yields preliminary solution $\vec{e}^{\,(i)}(\vec{x})$ after each iteration i
(with i coefficients = result of limited optimization process)
limited optimization feasible for small learning samples
(avoid over-adaptation)

Example

$\vec{x}$	r_k	$\vec{y}$	$P(R_1 \mid \vec{x})$	$e_1^{(1)}(\vec{x})$	$e_1^{(2)}(\vec{x})$	$e_1^{(3)}(\vec{x})$
(1,1)	R_1	(1,0)	0.67	0.6	0.60	0.67
(1,1)	R_1	(1,0)	0.67	0.6	0.60	0.67
(1,1)	R_2	(0,1)	0.67	0.6	0.60	0.67
(1,0)	R_1	(1,0)	0.50	0.6	0.60	0.50
(1,0)	R_2	(0,1)	0.50	0.6	0.60	0.50
(0,1)	R_1	(1,0)	0.33	0.0	0.33	0.33
(0,1)	R_2	(0,1)	0.33	0.0	0.33	0.33
(0,1)	R_2	(0,1)	0.33	0.0	0.33	0.33

Define $\vec{v}(\vec{x}) = (1, x_1, x_2)$

$$M = \frac{1}{8} \cdot \begin{pmatrix} 8 & 5 & 6 & 4 & 4 \\ 5 & 5 & 3 & 3 & 2 \\ 6 & 3 & 6 & 3 & 3 \end{pmatrix}.$$

$$\begin{aligned} d_1^{(1)} &= 0.50 \\ d_2^{(1)} &= 0.52 \\ d_3^{(1)} &= 0.50 \end{aligned}$$

$e_1^{(1)}(\vec{x}) = \frac{3}{5}x_1 \qquad e_2^{(1)} = \frac{2}{5}x_1$

$$M^{(2)} = \frac{1}{8} \cdot \begin{pmatrix} 3 & 0 & 3 & 1 & 2 \\ 5 & 5 & 3 & 3 & 2 \\ 3 & 0 & 4.2 & 1.2 & 1.8 \end{pmatrix}.$$

$$\begin{aligned} d_1^{(2)} &= 0.56 \\ d_3^{(2)} &= 0.27 \end{aligned}$$

$e_1^{(2)} = 0.33 + 0.27x_1 \qquad e_2^{(2)} = 0.67 - 0.27x_1.$

$$M^{(3)} = \frac{1}{8} \cdot \begin{pmatrix} 3 & 0 & 3 & 1 & 2 \\ 5 & 5 & 3 & 3 & 2 \\ 0 & 0 & 1.2 & 0.2 & -0.2 \end{pmatrix},$$

$e_1^{(3)} = 0.17 + 0.33x_1 + 0.17x_2 \qquad e_2^{(3)} = 0.83 - 0.33x_1 - 0.17x_2.$

$\vec{x}$	r_k	$\vec{y}$	$P(R_1\|\vec{x})$	$e_1^{(3)}(\vec{x})$	$e_1^{(3)'}(\vec{x})$
(1,1)	R_1	(1,0)	0.67	0.67	0.69
(1,1)	R_1	(1,0)	0.67	0.67	0.69
(1,1)	R_2	(0,1)	0.67	0.67	0.69
(1,0)	R_1	(1,0)	0.50	0.50	0.46
(1,0)	R_2	(0,1)	0.50	0.50	0.46
(0,1)	R_1	(1,0)	0.33	0.33	0.31
(0,1)	R_2	(0,1)	0.33	0.33	0.31
(0,1)	R_2	(0,1)	0.33	0.33	0.31
(0,0)	R_2	(0,1)	0.00	0.17	0.08

$$M' = \frac{1}{9} \cdot \begin{pmatrix} 9 & 5 & 6 & 4 & 5 \\ 5 & 5 & 3 & 3 & 2 \\ 6 & 3 & 6 & 3 & 3 \end{pmatrix} \qquad e_1^{(3)'} = 0.08 + 0.38x_1 + 0.23x_2$$

Experimental results

effectiveness measure: normalized recall R_{norm}
(for multivalued relevance scales or preference relation)

S^+ # doc. pairs in correct order

S^- # doc. pairs in wrong order

S^+_{max} max. # doc. pairs in correct order

$$R_{norm} = \frac{1}{2}\left(1 + \frac{S^+ - S^-}{S^+_{max}}\right).$$

random ordering of documents will have $R_{norm} = 0.5$ in the average

retr.fct.	sample	R^{μ}_{norm}	R^{M}_{norm}
e_1	B	0.752	0.754
e_1'		0.774*	0.751
e_2		0.752	0.763
e_2'		0.771*	0.745
e_1	C	0.721	0.753
e_1'		0.771*	0.740
e_2		0.721	0.714
e_2'		0.769*	0.710
cosine		0.668	0.728
e_1	A	0.741	0.775
e_1'		0.764*	0.756
e_2		0.741	0.771
e_2'		0.778*	0.760
cosine		0.727	0.769

retr. fct.	relev. scale	learn. sample	test sample	R^{μ}_{norm}	R^{M}_{norm}
e_1	$\mathcal{R}_2$	B	C	0.721	0.753
e_1	$\mathcal{R}_5$	B	C	0.721	0.749
e_1	$\mathcal{R}_2$	B	A	0.741	0.775
e_1	$\mathcal{R}_5$	B	A	0.749	0.772
e_2	$\mathcal{R}_2$	B	C	0.721	0.714
e_2	$\mathcal{R}_5$	B	C	0.720	0.707
e_2	$\mathcal{R}_5$	B	A	0.747	0.764
e_2	$\mathcal{R}_2$	B	A	0.741	0.771

Retrieval function based on binary ($\mathcal{R}_2$) and multi-valued relevance ($\mathcal{R}_5$) scales

retr. fct.	relev. scale	learn. sample	test sample	R^{μ}_{norm}	R^{M}_{norm}
e_1	$\mathcal{R}_2$	B/10	C	0.699	0.713
e_1	$\mathcal{R}_5$	B/10	C	0.695	0.718
e_1	$\mathcal{R}_2$	B/10	A	0.725	0.713
e_1	$\mathcal{R}_5$	B/10	A	0.730	0.732
e_2	$\mathcal{R}_2$	B/10	C	0.689	0.692
e_2	$\mathcal{R}_5$	B/10	C	0.702	0.689
e_2	$\mathcal{R}_5$	B/10	A	0.711	0.727
e_2	$\mathcal{R}_2$	B/10	A	0.696	0.716

Retrieval function based on binary ($\mathcal{R}_2$) and multi-valued relevance ($\mathcal{R}_5$) scales adapted on a small (B/10 = every 10th request-document pair of B) learning sample

Model-oriented vs. description-oriented approaches

Model-oriented approaches:

- refer to specific representation
- based on certain explicit assumptions
- strict theoretical model
- quality depends on validity of assumptions

Description-oriented approach:

- flexible w.r.t. representation
- most assumptions implicit
- heuristic definition of relevance description
- better adaptation to the application data

The BII model

(Binary independence indexing)
regards one document in relation to a number of queries

q_k query representation = set of terms $q_k^T \subset T$
binary vector $\vec{z}_k = (z_{k_1}, \ldots, z_{k_n}) =$ with

$$z_{k_i} = \begin{cases} 1, & \text{if } t_i \in q_k^T \\ 0, & \text{otherwise} \end{cases}$$

d_m document representation: not further specified

d_m^T terms with weights w.r.t. the document.

$P(R|q_k, d_m)$ $= P(R|\vec{z}_k, d_m)$: probability that document with representation d_m will be judged relevant w.r.t. a query with representation $q_k = q_k^T$

apply Bayes' theorem:

$$P(R|\vec{z}_k, d_m) = P(R|d_m) \cdot \frac{P(\vec{z}_k|R, d_m)}{P(\vec{z}_k|d_m)} \tag{3}$$

$P(R|d_m)$ prob. that d_m will be judged relevant to arbitrary request

$P(\vec{z}_k|d_m)$ prob. of query with rep. $\vec{z}_k$

independence assumption:
independent distribution of terms in all queries to which a document with representation d_m is relevant:

$$P(\vec{z}_k|R, d_m) = \prod_{i=1}^{n} P(z_{k_i}|R, d_m)$$

$$
\begin{aligned}
P(R|\vec{z}_k, d_m) &= \frac{P(R|d_m)}{P(\vec{z}_k|d_m)} \cdot \prod_{i=1}^{n} P(z_{k_i}|R, d_m) \\
&= \frac{P(R|d_m)}{P(\vec{z}_k|d_m)} \cdot \prod_{i=1}^{n} \frac{P(R|z_{k_i}, d_m)}{P(R|d_m)} \cdot P(z_{k_i}|d_m)
\end{aligned}
$$

$P(\vec{z}_k|d_m)$ and $P(z_{k_i}|d_m)$ are independent of a specific document (since we always regard all documents w.r.t. a query):

$$
\begin{aligned}
P(R|\vec{z}_k, d_m) &= \frac{\prod_{i=1}^{n} P(z_{k_i})}{P(\vec{z}_k)} \cdot P(R|d_m) \cdot \prod_{i=1}^{n} \frac{P(R|z_{k_i}, d_m)}{P(R|d_m)} \\
&= \frac{\prod_{i=1}^{n} P(z_{k_i})}{P(\vec{z}_k)} \cdot P(R|d_m) \cdot \prod_{z_{k_i}=1} \frac{P(R|z_{k_i} = 1, d_m)}{P(R|d_m)} \cdot \\
&\quad \cdot \prod_{z_{k_i}=0} \frac{P(R|z_{k_i} = 0, d_m)}{P(R|d_m)} \qquad (4)
\end{aligned}
$$

additional simplifying assumption:
relevance of a doc. d_m with respect to a query q_k depends only on the terms from q_k^T, and not on other terms

$$\prod_{z_{k_i}=0} \frac{P(R|z_{k_i}=0, d_m)}{P(R|d_m)} = 1$$

constant factor c_k for a given query q_k, not required for ranking:

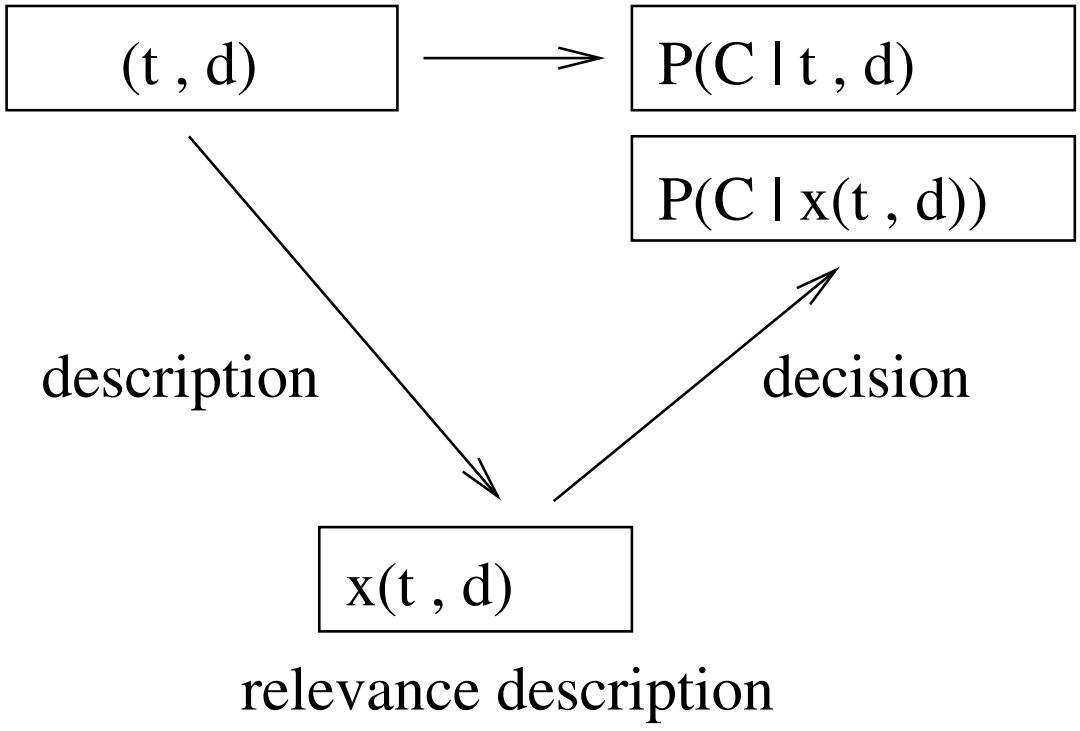
$$\frac{\prod_{i=1}^{n} P(z_{k_i})}{P(\vec{z}_k)} = c_k$$

$P(R|z_{k_i} = 1, d_m) = P(R|t_i, d_m)$:
probabilistic index term weight of t_i w.r.t. d_m =
prob. that doc. d_m will be judged relevant to an arbitrary query, containing t_i.
d_m^T should contain at least those terms from T for which $P(R|t_i, d_m) \neq P(R|d_m)$.
assume that $P(R|t_i, d_m) = P(R|d_m)$ for all $t_i \notin d_m^T$:

$$P(R|q_k, d_m) = c_k \cdot P(R|d_m) \cdot \prod_{t_i \in q_k^T \cap d_m^T} \frac{P(R|t_i, d_m)}{P(R|d_m)}. \qquad (5)$$

application of BII model in this form nearly impossible, because of lack of relevance information for specific term-document pairs.

A description-oriented indexing approach
Darmstadt Indexing Approach



regard features of terms in documents instead of the document-term pairs itself
(description-related learning strategy)

Description step

relevance description $x(t_i, d_m)$ contains attribute values of

- the term t_i
- the document d_m
- the term-document relationship

approach makes no additional assumptions about the choice of the attributes or the structure of x
$\longrightarrow$ actual definition of relevance descriptions can be adapted to the specific application context
(representation of documents, amount of learning data available)

Decision step

instead of $P(R|t_i, d_m)$,
estimate $P(R|x(t_i, d_m))$
$P(R|t_i, d_m)$:
regard a single document d_m with respect to all queries containing t_i
$P(R|x(t_i, d_m))$:
regard set of all term-document pairs with the same relevance description x
learning example $L \subset \underline{Q} \times \underline{D} \times \mathcal{R}$
(query-document with relevance judgements)
$L = \{(\underline{q}_k, \underline{d}_m, r_{km})\}$.
form relevance descriptions for the terms common to query and document :
bag of relevance descriptions with relevance judgements
$L^x = [(x(t_i, d_m), r_{km}) | t_i \in q_k^T \cap d_m^T \wedge (\underline{q}_k, \underline{d}_m, r_{km}) \in L]$.
estimate parameters $P(R|x(t_i, d_m))$ by applying probabilistic classification procedures (e.g. LSP) $\rightarrow$ indexing function $e(x(t_i, d_m))$

Example for description-oriented approach

query	doc.	judg.	term	$\vec{x}$
q_1	d_1	R	t_1	$(1,1)$
			t_2	$(0,1)$
			t_3	$(1,2)$
q_1	d_2	$\bar{R}$	t_1	$(0,2)$
			t_3	$(1,1)$
			t_4	$(0,1)$
q_2	d_1	R	t_2	$(0,2)$
			t_5	$(0,2)$
			t_6	$(1,1)$
			t_7	$(1,2)$
q_2	d_3	$\bar{R}$	t_5	$(0,1)$
			t_7	$(0,1)$

	$P(R\|\vec{x})$	
$\vec{x}$	E_X	E_{BII}
$(0,1)$	$1/4$	$1/3$
$(0,2)$	$2/3$	$1/2$
$(1,1)$	$2/3$	$2/3$
$(1,2)$	1	1

2 different event spaces:
E_{BII}: equiprobable query-document pairs
E_X: equiprobable relevance descriptions

Indexing function

Data used (same as in SMART):

tf_{mi}: within-document frequency (wdf) of t_i in d_m.

$max\ tf_m$: maximum wdf tf_{mi} of all terms $t_i \in d_m^T$.

n_i: number of documents in which t_i occurs.

$|\underline{D}|$: number of documents in the collection.

$|d_m^T|$: number of different terms in d_m.

ta_{mi}: $= 1$, if t_i occurs in the title of d_m, and 0 otherwise.

Relevance description elements:

$$
\begin{aligned}
x_1 &= tf_{mi} \\
x_2 &= 1/max\ tf_m \\
x_3 &= \log(n_i/|D|) \\
x_4 &= \log|d_m^T| \\
x_5 &= ta_{mi}
\end{aligned}
$$

Indexing functions:

$$e_L = a_0 + a_1\ tf_{mi} + a_2/max\ tf_m + a_3\ \log(n_i/|\underline{D}|) + a_4\ \log|d_m^T|,$$

$$e_{ta} = a_0 + a_1\ tf_{mi} + a_2/max\ tf_m + a_3\ \log(n_i/|\underline{D}|) + a_4\ \log|d_m^T| + a_5 ta_{mi}$$

Retrieval functions

q_k^T — set of query terms

d_m^T — set of document terms

u_{mi} — indexing weight $e(\vec{x}(t_i, d_m))$

c_{ki} — query term weight (see below)

utility–theoretic retrieval function [Wong & Yao 89]:

$$\varrho(q_k, d_m) = \sum_{t_i \in q_k^T \cap d_m^T} c_{ki} \cdot u_{mi}$$

- ϱ_{bin} binary query term weights ($c_{ki} = 1$ for all $t_i \in q_k^T$)
- ϱ_{tf} c_{ki} = within–query frequency of t_i in q_k
- ϱ_{tfidf} c_{ki} = *tfidf* with within–query frequencies

Experimental results

collection	$tf \times idf$	e_L			e_{ta}	
	ϱ_{tfidf}	ϱ_{bin}	ϱ_{tf}	ϱ_{tfidf}	ϱ_{bin}	ϱ_{tf}
CACM	0.2963	0.3046 + 2.8%	0.3371 + 13.8%	0.3283 + 10.8%	0.3148 + 6.2%	0.3464 + 16.9%
CISI	0.2099	0.1731 - 17.5%	0.2288 + 9.0%	0.2052 - 2.2%	0.1751 - 16.6%	0.2311 + 10.1%
CRAN	0.3816	0.4265 + 11.8%	0.4293 + 12.5%	0.3922 + 2.8%	0.4554 + 19.3%	0.4556 + 19.4%
INSPEC	0.2489	0.2307 - 7.3%	0.2708 + 8.8%	0.2502 + 0.5%	0.2326 - 6.5%	0.2694 + 8.2%
NPL	0.2138	0.2834 + 32.6%	0.2834 + 32.6%	0.2382 + 11.4%	- -	- -

Comparison of different retrieval functions (test sample, top, E_x)

average precision at three recall points (0.25, 0.50, 0.75)

- probabilistic indexing with ϱ_{tf} better than SMART approach
- SMART approach works without relevance information
- probabilistic indexing can be extended easily to more complex representations

Retrieval with probabilistic indexing

probabilistic model for ranking of documents with probabilistic indexing weights
probabilistic indexing:
assume a fixed number of binary indexings per document
$\longrightarrow$ extended event space $\underline{Q} \times \underline{D} \times I$:

- $\underline{Q}$ set of queries
- $\underline{D}$ set of documents
- I set of indexers

single event: query-document pair with

- relevance judgement
- specific (binary) indexing
- relevance descriptions for the terms w.r.t. the document

Representations and descriptions:

- query representation $q_k = (q_k^T, q_k^J)$:
  (like in BIR model) set of query terms q_k^T +
  set of relevance judgements $q_k^J = \{(\underline{d_m}, r(\underline{d_m}, \underline{q_k}))\}$
- query description q_k^D:
  set of query terms with associated weights
- document representation: $d_m = (\vec{d}_m, \vec{c}_m)$

  $\vec{d}_m = (d_{m_1}, \ldots, d_{m_n})$, where d_{m_i} is the relevance description of t_i w.r.t. d_m

  $\vec{c}_m = (c_{m_1}, \ldots, c_{m_n})$ with

$$c_{m_i} = \begin{cases} C_i, & \text{if } t_i \text{ has been assigned to } d_m \\ \bar{C}_i, & \text{otherwise} \end{cases}$$

- document description d_m^D:
  set of terms with indexing weights

$P(R|q_k, \vec{x})$: probability that document with relevance descriptions $\vec{x}$ is relevant w.r.t. q_k
apply Bayes' theorem:

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

linked dependence assumption:

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

yields

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

Assumptions:

- relevance description x_i of a term t_i depends only on the correctness of t_i, independent of the correctness of other terms and relevance
- correctness of a term (w.r.t. a document) depends only on relevance, independent of the correctness of other terms

$$O(R|q_k,\vec{x}) =$$

$$O(R|q_k)\prod_{i=1}^{n}\frac{P(x_i|C_i)\cdot P(C_i|R,q_k)+P(x_i|\bar{C}_i)\cdot P(\bar{C}_i|R,q_k)}{P(x_i|C_i)\cdot P(C_i|\bar{R},q_k)+P(x_i|\bar{C}_i)\cdot P(\bar{C}_i|\bar{R},q_k)}$$

$$= O(R|q_k)\prod_{i=1}^{n}\frac{\frac{P(C_i|x_i)}{P(C_i)}\cdot P(C_i|R,q_k)+\frac{P(\bar{C}_i|x_i)}{P(\bar{C}_i)}\cdot P(\bar{C}_i|R,q_k)}{\frac{P(C_i|x_i)}{P(C_i)}\cdot P(C_i|\bar{R},q_k)+\frac{P(\bar{C}_i|x_i)}{P(\bar{C}_i)}\cdot P(\bar{C}_i|\bar{R},q_k)}$$

u_{m_i} $= P(C_i|x_i{=}d_{m_i})$: probabilistic indexing weight of t_i w.r.t. d_m
= probability that arbitrary indexer assigned t_i to d_m

q_i $= P(C_i)$: probability that arbitrary indexer assigned t_i to arbitrary document
= average indexing weight of t_i in the collection

p_{ik} $= P(C_i|R,q_k)$: probability that arbitrary indexer assigned t_i to arbitrary document relevant to q_k
= average indexing weight of t_i in relevant documents

r_{ik} $= P(C_i|\bar{R},q_k)$: probability that arbitrary indexer assigned t_i to arbitrary doc. nonrelevant to q_k = average indexing weight of t_i in nonrelevant docs

assume that $P(C_i|R, q_k) = P(C_i|\bar{R}, q_k)$ for all $t_i \notin q_k^T$

$$O(R|q_k, \vec{x}=\vec{d}_m) = O(R|q_k) \prod_{t_i \in q_k^T} \frac{\frac{u_{m_i}}{q_i} p_{ik} + \frac{1-u_{m_i}}{1-q_i}(1-p_{ik})}{\frac{u_{m_i}}{q_i} r_{ik} + \frac{1-u_{m_i}}{1-q_i}(1-r_{ik})}$$

approximation for $P(C_i|\bar{R}, q_k) \approx P(C_i)$:

$$O(R|q_k, \vec{x}=\vec{d}_m) \approx O(R|q_k) \prod_{t_i \in q_k^T} \frac{u_{m_i}}{q_i} p_{ik} + \frac{1-u_{m_i}}{1-q_i}(1-p_{ik})$$

Parameter estimation

$\underline{D}_k^R$ — set of docs. judged relevant w.r.t. q_k

$$q_i = \frac{1}{|\underline{D}|} \sum_{\underline{d}_j \epsilon \underline{D}} u_{j_i}$$

$$p_{ik} = \frac{1}{|\underline{D}_k^R|} \sum_{\underline{d}_j \epsilon \underline{D}_k^R} u_{j_i}$$

Parameter estimation

parameter estimation affects retrieval quality observed in experiments!

2 problems:

1. estimation sample selection
   (random sample vs. top ranking documents)
   - estimate parameters for nonrelevant documents from all documents not known to be relevant
   - description-oriented approaches mostly assume a sample representative for the documents to be ranked
     (and not representative for the whole collection)
2. estimation method (data $\longrightarrow$ parameters)

Estimation methods

collection of documents with features e_i
estimation of $P(e_i|e_j)$
BIR model:
e_j = *relevance / non-relevance w.r.t.* q_k
e_i = *presence / absence of a term* t_i

- g number of documents in (random) sample, where
- f documents with feature e_j and
- h documents with e_i and e_j

Task:
computation of estimate $p(e_i|e_j, (h, f, g))$ for $P(e_i|e_j)$, given (h, f, g)
maximum likelihood estimate: $p(e_i|e_j, (h, f, g)) = h/f$

- not defined for $f = h = 0$
- biased estimate

Bayesian probability estimation

Q parameter to be estimated (continuous random variable)
$f(q)$ prior distribution of parameter Q
X discrete random variable with values $x_1, x_2, \ldots$
$P(X{=}x_k|Q{=}q)$: probability that X will take the value x_k, given that Q has the value q
posterior distribution of q:

$$g(q|x_k) = \frac{f(q) \cdot P(X{=}x_k|Q{=}q)}{\int_{-\infty}^{\infty} f(q) \cdot P(X{=}x_k|Q{=}q)dq}$$

estimate for q requires application of further methods, e.g. cost function

Estimates for beta prior

assume beta distribution as prior distribution:

$$f(p) = \frac{1}{B(a,b)}\, p^{a-1}(1-p)^{b-1}$$

with $B(a,b)=\Gamma(a)\cdot\Gamma(b)/\Gamma(a+b)$
$a, b > 0$: parameters to be chosen
application with (f, h) observed:

$$g(p|\tfrac{h}{f}) = \frac{p^{a-1}(1-p)^{b-1}\binom{f}{h}p^h(1-p)^{f-h}}{\int_0^1 p^{a-1}(1-p)^{(b-1)}\binom{f}{h}p^h(1-p)^{f-h}dp}$$

$B(a,b) = \int_0^1 p^{a-1}(1-p)^{b-1}dp$ yields
posterior distribution:

$$g(p|\tfrac{h}{f}) = \frac{p^{h+a-1}(1-p)^{f-h+b-1}}{B(h+a, f-h+b)}$$

apply loss function

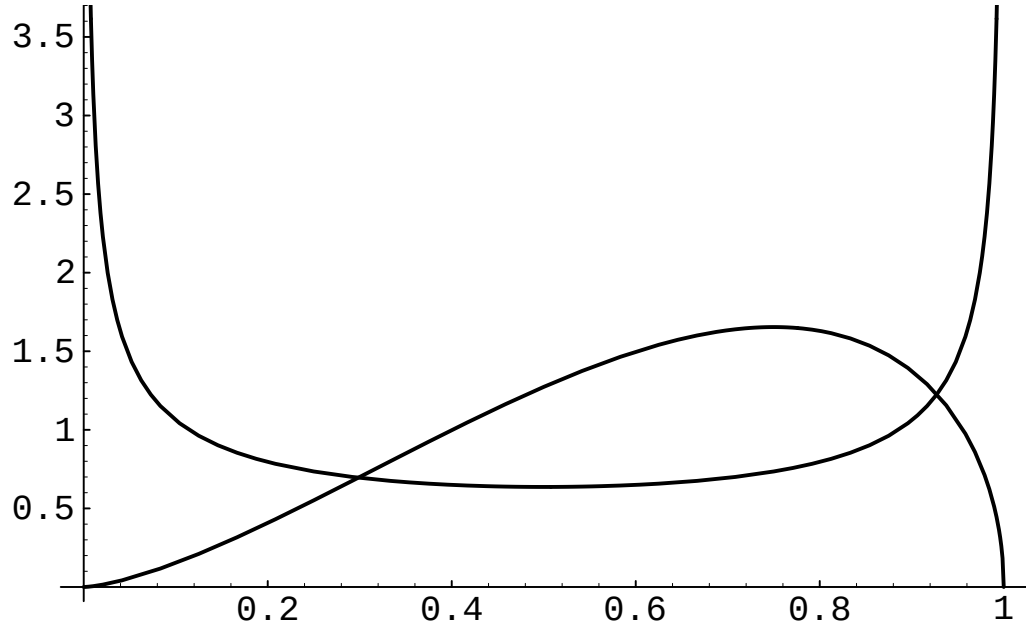
$$L(\hat{p}, p_{ij}) = (\hat{p} - p_{ij})^2$$

seek for estimate p_L minimizing the expectation of the loss function:

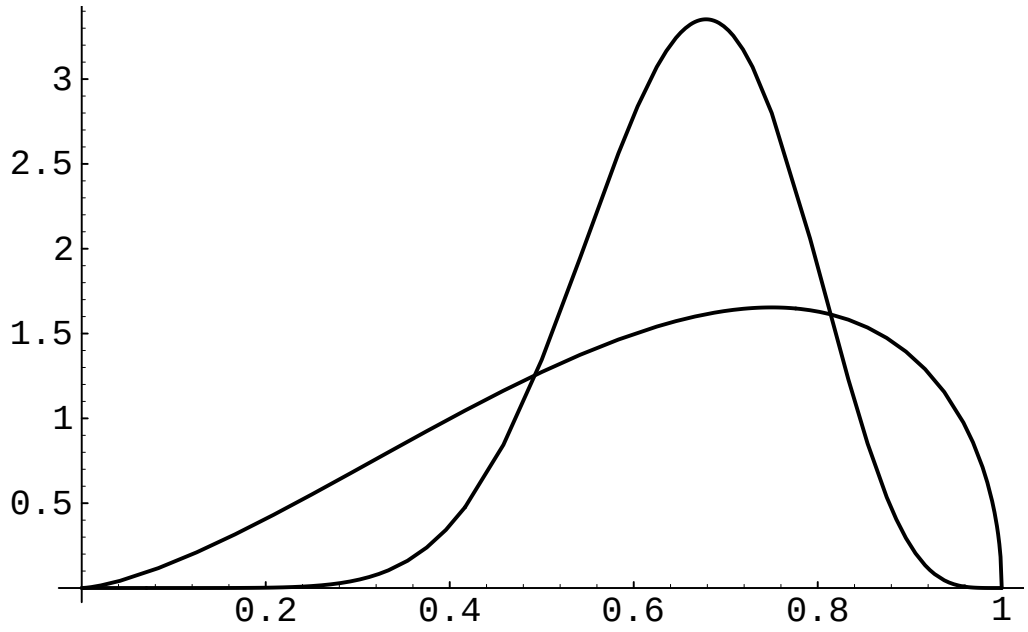
$$\frac{d}{dp_L} \int_0^1 L(p, p_L) g(p) dp \stackrel{!}{=} 0$$

yields

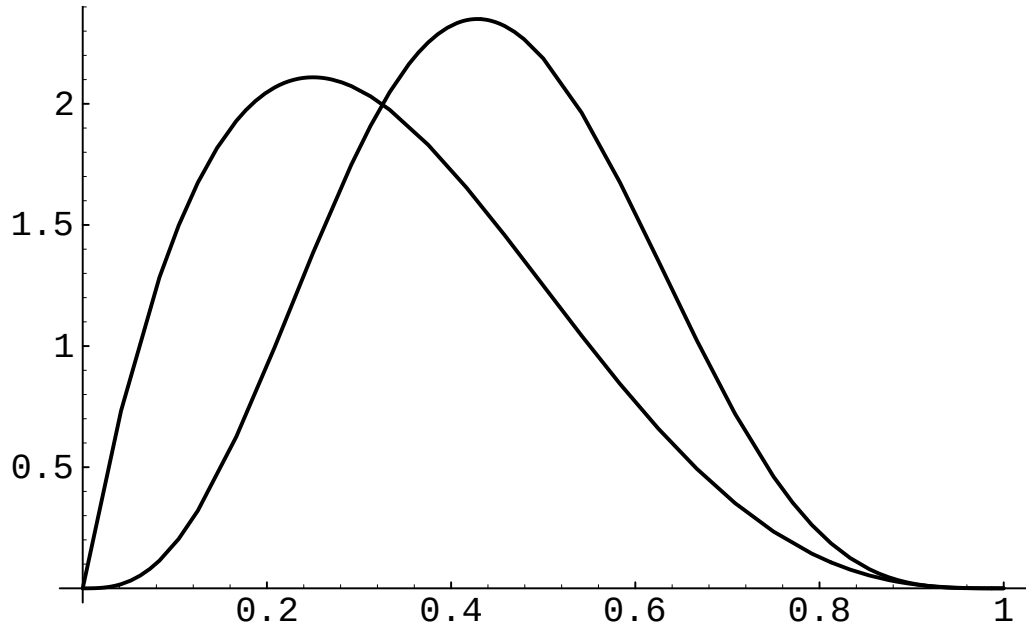
$$p_L = \frac{h + a}{f + a + b}$$



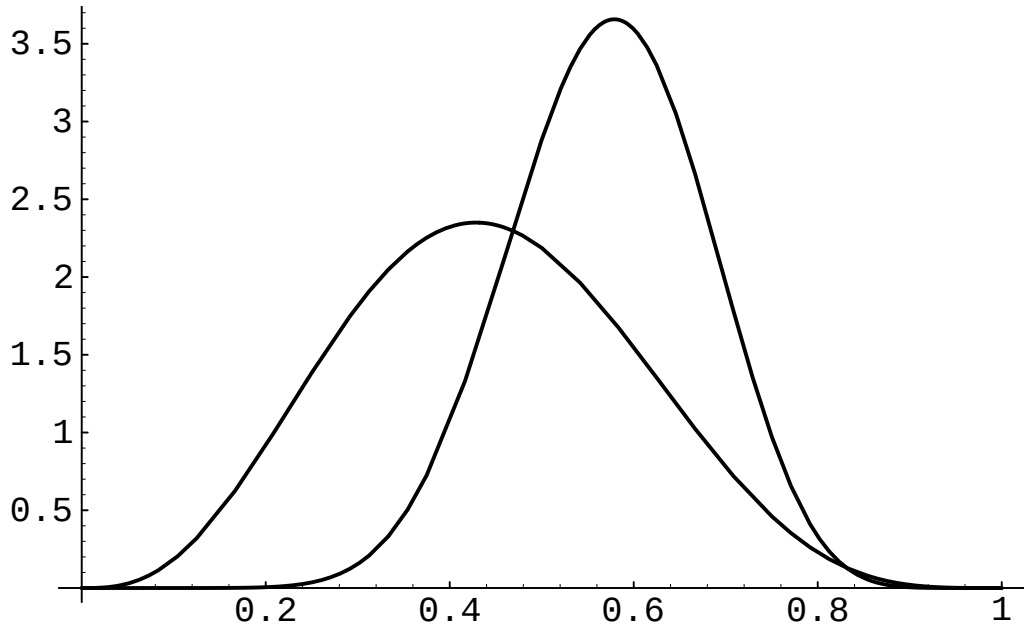
Prior (upper curve) and posterior distribution for $a = b = 0.5$, $f = 3$, $h = 2$



Posterior distributions for $a = b = 0.5$, $f = 3$, $h = 2$ and $f' = 15$, $h' = 10$



Prior (left curve) and posterior distribution for $a = 2$, $b = 4$, $f = 3$, $h = 2$



Posterior distributions for $a = 2$, $b = 4$, $f = 3$, $h = 2$ and $f' = 15$, $h' = 10$

Experimental results

f	$Z_{g'}(h, f)$	a	b
4	139	-3.42	0.47
5	341	2.64	5.85
6	349	-1.54	1.48
7	546	-1.61	1.33
8	526	-1.54	1.76
9	816	-1.29	1.84
10	374	-1.46	1.31

parameters of the beta distribution. derived from a sample of 1 000 PHYS documents.

$Z_{g'}$: # pairs (e_i, e_j) in the learning sample ($g = 78\,000$).

sample	$Z_{g'}$	estimate	s^2
X	2852	p_{ML}	0.271
		p_{L_1}	0.231
		p_{opt}	0.231
Z	2709	p_{ML}	0.265
		p_{L_1}	0.226
		p_{opt}	0.224

different kinds of estimates for $f = 4 \ldots 10, h \geq 4$ and $\frac{h}{f} > 0.4$

$Z_{g'}$: # pairs (e_i, e_j)

s^2: average quadratic error