6 Probabilistic Retrieval Models

- Notations
- Binary Independence Retrieval model
- Probability Ranking Principle
6.1 Notations

\[ Q \xrightarrow{\alpha_Q} Q \xrightarrow{\beta_Q} Q^D \]

\[ D \xrightarrow{\alpha_D} D \xrightarrow{\beta_D} D^D \]

\[ \mathcal{R} \xrightarrow{rel.} Q \xrightarrow{\varrho} IR \]

\[ \mathcal{R} \xrightarrow{judg.} D \xrightarrow{\varrho} D^D \]

\[ q \in Q \quad query \]
\[ q_k \in Q \quad query \ representation \]
\[ q_k^D \in Q^D \quad query \ description \]
\[ \mathcal{R} \quad relevance \ scale \]
\[ \varrho \quad retrieval \ function \]

\[ d \in D \quad document \]
\[ d_m \in D \quad document \ representation \]
\[ d_m^D \in D^D \quad document \ description \]
6.2 Binary independence retrieval model

6.2.1 Retrieval functions for binary indexing

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

\[ q_k \in Q: \text{query representation} \quad q_k^T: \text{set of query terms} \]

\[ d_m \in D: \text{document representation} \quad d_m^T: \text{set of document terms} \]

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

\[ \varrho_{BIR}(q_k, d_m) = \sum_{t_i \in q_k^T \cap d_m^T} c_{ik} \]
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 $\vec{x} = (x_1, \ldots, x_n)$ 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 $q_k$ (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:

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

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

\[
q_{ik} = P(x_i=1|\bar{R}, q_k) \text{ prob. that } t_i \text{ occurs in arbitrary nonrelevant doc.}
\]
assume that $p_{ik} = q_{ik}$ for all $t_i \notin q_k^T$

\[
O(R|q_k, d_m^T) = O(R|q_k) \prod_{t_i \in d_m^T \cap q_k^T} \frac{p_{ik}}{q_{ik}} \cdot \prod_{t_i \in q_k^T \setminus d_m^T} \frac{1 - p_{ik}}{1 - q_{ik}}
\]

\[
= O(R|q_k) \prod_{t_i \in d_m^T \cap q_k^T} \frac{p_{ik}(1 - q_{ik})}{q_{ik}(1 - p_{ik})} \cdot \prod_{t_i \in q_k^T} \frac{1 - p_{ik}}{1 - q_{ik}}
\]

Only first product varies for different documents with respect to the same request $q_k$ → regard only this product for ranking
use logarithm:
\[ c_{ik} = \log \frac{p_{ik}(1 - q_{ik})}{q_{ik}(1 - p_{ik})} \]

retrieval function:
\[ \varrho_{BIR}(q_k, d_m) = \sum_{t_i \in d_m \cap q_k^T} c_{ik} \]
6.2.3 Application of the BIR model

Parameter estimation for $q_{ik}$

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

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

$$c_{ik} = \log \frac{p}{1-p} + \log \frac{1-q_{ik}}{q_{ik}}$$

$$= c_p + \log \frac{N-n_i}{n_i}$$

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

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

\[ r: \ # \text{documents judged relevant for request } q_k \]

\[ r_i: \ # \text{relevant documents with term } t_i \]

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

improved estimates (see parameter estimation methods):

\[ p_{ik} \approx \frac{r_i + 0.5}{r + 1} \]
## BIR example

| $d_m$ | $x_1$ | $x_2$ | $P(R|\vec{x})$ | $r(d_m)$ | BIR | $P(R|\vec{x})$ | BIR |
|-------|-------|-------|-----------------|----------|-----|-----------------|-----|
| $d_1$ | 1     | 1     | 0.48            |          |     |                 |     |
| $d_2$ | 1     | 1     | 0.76            |          |     |                 |     |
| $d_3$ | 1     | 1     | 0.80            |          |     |                 |     |
| $d_4$ | 1     | 1     |                 |          |     |                 |     |
| $d_5$ | 1     | 1     |                 |          |     |                 |     |
| $d_6$ | 1     | 1     |                 |          |     |                 |     |
| $d_7$ | 1     | 1     |                 |          |     |                 |     |
| $d_8$ | 1     | 1     |                 |          |     |                 |     |
| $d_9$ | 1     | 1     |                 |          |     |                 |     |
| $d_{10}$ | 1     | 1     |                 |          |     |                 |     |
| $d_{11}$ | 1     | 1     |                 |          |     |                 |     |
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 \( d_j \):

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

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 \( i \)

\[
EC(q, l) = EC(q, d_{r(1)}, d_{r(2)}, \ldots, d_{r(l)})
\]

\[
= \sum_{i=1}^{l} EC(q, d_{r(i)})
\]

Minimum total costs → minimize \( \sum_{i=1}^{l} EC(q, d_{r(i)}) \) →

\( r(i) \) should order documents by ascending costs

Norbert Fuhr
**decision-theoretic rule:**

\[
EC(q, d_{r(i)}) \leq EC(q, d_{r(i+1)}) \iff \\
C \cdot P(R|q, d_{r(i)}) + \bar{C}(1 - P(R|q, d_{r(i)})) \leq C \cdot P(R|q, d_{r(i+1)}) + \bar{C}(1 - P(R|q, d_{r(i+1)}))
\]

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

**rank documents according to their decreasing probability of relevance!**
6.3.2 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)

\[
O(a|b) = \frac{P(b|a)P(a)}{P(b|\bar{a})P(\bar{a})}
\]
\[
\log O(a|b) = \log \frac{P(b|a)}{P(b|\bar{a})} + \log O(a)
\]
\[
\logit P(a|b) = \log \frac{P(b|a)}{P(b|\bar{a})} + \logit P(a)
\]

with $\logit P(x) = \log O(x)$
\[ \rho = P(\text{doc.retrieved}|\text{doc.relevant}) \]
\[ \phi = P(\text{doc.retrieved}|\text{doc.nonrelevant}) \]
\[ \pi = P(\text{doc.relevant}|\text{doc.retrieved}) \]
\[ \gamma = P(\text{doc.relevant}) \]
\[ \rho(d_i) = P(\text{doc.is } d_i|\text{doc.relevant}) \]
\[ \phi(d_i) = P(\text{doc.is } d_i|\text{doc.nonrelevant}) \]
\[ \pi(d_i) = P(\text{doc.relevant}|\text{doc.is } d_i) \text{ (probability of relevance)} \]

\[ S: \text{ set of retrieved documents} \]
\[ \rho = \sum_{d_i \in S} \rho(d_i) \]
\[ \phi = \sum_{d_i \in S} \phi(d_i) \]

\[ \logit \pi(d_i) = \log \frac{\rho(d_i)}{\phi(d_i)} + \logit \gamma \]
\[ \rho(d_i) = x_i \cdot \phi(d_i) \quad \text{with} \]
\[ x_i = \exp(\logit \pi(d_i) - \logit \gamma) \]
1. cutoff defined by $\phi$ (fallout)

$$\phi = \sum_{d_i \in S} \phi(d_i)$$

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

$\leadsto$ maximize $\rho$ (recall) by including docs with highest values of $\pi(d_i)$

$\hat{=} \text{rank according to prob. of relevance}$

2. given # documents retrieved

$\leadsto$ maximize expected recall, minimize expected fallout

3. cutoff defined by $\rho$ (recall)

$\leadsto$ minimize fallout
\[
\text{logit } \pi = \log(\rho/\phi) + \text{logit } \gamma
\]

4. expected precision is maximized for given recall / fallout / \# documents retrieved
6.3.3 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 at computing 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 General concepts of probabilistic IR models

6.4.1 Conceptual model

\[ \mathcal{R} \xleftarrow{\text{rel.}} Q \xrightarrow{\alpha_Q} Q \xrightarrow{\beta_Q} Q^D \]

\[ \xrightarrow{\varrho} D \xrightarrow{\alpha_D} D \xrightarrow{\beta_D} D^D \]

\[ \xrightarrow{\varnothing} IR \]
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 $q_k^J = \{(d_m, r(d_m, q_k))\}$

- 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 $d_m^D = \text{document representation } d_m^T$
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)
6.4.2 Parameter learning in IR

Learning approaches in IR

Norbert Fuhr
6.4.3 Event space

Event space: $Q \times D$

single element: query-document pair $(q_k, d_m)$

all elements are equiprobable

relevance judgement $(q_k, d_m) \in \mathcal{R}$

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 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
6.5 Relevance models

6.5.1 Optimum polynomial retrieval functions

Basic concepts
regard retrieval as (probabilistic) classification task
(classify objects into one of \( n \) classes)

objects: query-document pairs \((q_k, d_m)\)
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:

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

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

\( \vec{e}_{opt}(\vec{x}) \) yields probabilistic estimates \( P(R_l|\vec{x}) \) in the components of \( \hat{\vec{y}} \)

variation problem cannot be solved in its general form

→ restrict search to predefined class of functions

general variation problem → parameter optimization task.
resulting functions yield least squares approximations of $\vec{e}_{opt}$:

approximation with respect to the expression

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

yields the same result as an optimization fulfilling the condition

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

→ parameter optimization 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

\[ \vec{v}(\vec{x}) = (1, x_1, x_2, \ldots, x_N, x_1^2, x_1x_2, \ldots) \]

\( N = \) number of dimensions of \( \vec{x} \)

class of polynomials is given by the components

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

(mostly linear and quadratic polynomials regarded)
regression function:

\[ \vec{e}(\vec{x}) = A^T \cdot \vec{v}(\vec{x}) \]

where \( A = (a_{i,l}) \) with \( i=1, \ldots, L; l=1, \ldots, n \) is the coefficient matrix

\( P(R_l|\vec{x}) \) is approximated by the polynomial

\[ e_l(\vec{x}) = a_{1,l} + a_{2,l} \cdot x_1 + a_{3,l} \cdot x_2 + \ldots + a_{N+1,l} \cdot x_N + \\
+ a_{N+2,l} \cdot x_1^2 + a_{N+3,l} \cdot x_1 \cdot x_2 + \ldots \]

coefficient matrix is computed by solving the equation system

\[ E(\vec{v} \cdot \vec{v}^T) \cdot A = E(\vec{v} \cdot \vec{y}^T). \quad (1) \]
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 \(M\):
     \[
     M = (\vec{u} \cdot \vec{u}^T \; \vec{u} \cdot \vec{y}^T).
     \]
     \((M\) 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 $i$th elimination step:
$M^{(i)} = (m^{(i)}_{lj})$ with $l=1, \ldots, n; j=1, \ldots, L + n$
reduction $d^{(i)}_j$ achieved by choosing component $j$:

$$
\begin{align*}
    d^{(i)}_j &= \frac{1}{m^{(i)}_{jj}} \cdot \sum_{l=L+1}^{L+n} m^{(i)}_{jl}^2.
\end{align*}
$$

(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

| $\vec{x}$ | $r_k$ | $\vec{y}$ | $P(R_1 | \vec{x})$ | $e_1^{(1)}(\vec{x})$ | $e_1^{(2)}(\vec{x})$ | $e_1^{(3)}(\vec{x})$ |
|---|---|---|---|---|---|---|
| (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 $\vec{v}(\vec{x}) = (1, x_1, x_2)$
\[
M = \frac{1}{8} \cdot \begin{pmatrix}
8 & 5 & 6 & 4 & 4 \\
5 & 5 & 3 & 3 & 2 \\
6 & 3 & 6 & 3 & 3 \\
\end{pmatrix}.
\]

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

\[
e^{(1)}_1(x) = \frac{3}{5} x_1 \quad e^{(1)}_2 = \frac{2}{5} x_1
\]

\[
d^{(1)}_1 = 0.50 \quad d^{(1)}_2 = 0.52 \quad d^{(1)}_3 = 0.50
\]

\[
e^{(2)}_1 = 0.33 + 0.27 x_1 \quad e^{(2)}_2 = 0.67 - 0.27 x_1.
\]

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

\[
e^{(3)}_1 = 0.17 + 0.33 x_1 + 0.17 x_2 \quad e^{(3)}_2 = 0.83 - 0.33 x_1 - 0.17 x_2.
\]

Norbert Fuhr
| $\vec{x}$ | $r_k$ | $\vec{y}$ | $P(R_1|\vec{x})$ | $e_1^{(3)}(\vec{x})$ | $e_1^{(3)'}(\vec{x})$ |
|---|---|---|---|---|---|
| (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 |

$$M' = \frac{1}{9} \cdot \begin{pmatrix} 9 & 5 & 6 & 4 & 5 \\ 5 & 5 & 3 & 3 & 2 \\ 6 & 3 & 6 & 3 & 3 \end{pmatrix}$$

$$e_1^{(3)'} = 0.08 + 0.38x_1 + 0.23x_2$$
Experimental results

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

$S^+ \#$ doc. pairs in correct order

$S^- \#$ doc. pairs in wrong order

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

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

random ordering of documents will have $R_{\text{norm}} = 0.5$ in the average
<table>
<thead>
<tr>
<th>retr.fct.</th>
<th>sample</th>
<th>$R^\mu_{norm}$</th>
<th>$R^M_{norm}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$e_1$</td>
<td>B</td>
<td>0.752</td>
<td>0.754</td>
</tr>
<tr>
<td>$e'_1$</td>
<td></td>
<td>0.774*</td>
<td>0.751</td>
</tr>
<tr>
<td>$e_2$</td>
<td></td>
<td>0.752</td>
<td>0.763</td>
</tr>
<tr>
<td>$e'_2$</td>
<td></td>
<td>0.771*</td>
<td>0.745</td>
</tr>
<tr>
<td>$e_1$</td>
<td>C</td>
<td>0.721</td>
<td>0.753</td>
</tr>
<tr>
<td>$e'_1$</td>
<td></td>
<td>0.771*</td>
<td>0.740</td>
</tr>
<tr>
<td>$e_2$</td>
<td></td>
<td>0.721</td>
<td>0.714</td>
</tr>
<tr>
<td>$e'_2$</td>
<td></td>
<td>0.769*</td>
<td>0.710</td>
</tr>
<tr>
<td>cosine</td>
<td></td>
<td>0.668</td>
<td>0.728</td>
</tr>
<tr>
<td>$e_1$</td>
<td>A</td>
<td>0.741</td>
<td>0.775</td>
</tr>
<tr>
<td>$e'_1$</td>
<td></td>
<td>0.764*</td>
<td>0.756</td>
</tr>
<tr>
<td>$e_2$</td>
<td></td>
<td>0.741</td>
<td>0.771</td>
</tr>
<tr>
<td>$e'_2$</td>
<td></td>
<td>0.778*</td>
<td>0.760</td>
</tr>
<tr>
<td>cosine</td>
<td></td>
<td>0.727</td>
<td>0.769</td>
</tr>
<tr>
<td>retr. fct.</td>
<td>relev. scale</td>
<td>learn. sample</td>
<td>test sample</td>
</tr>
<tr>
<td>-----------</td>
<td>-------------</td>
<td>---------------</td>
<td>-------------</td>
</tr>
<tr>
<td>$e_1$</td>
<td>$\mathcal{R}_2$</td>
<td>B</td>
<td>C</td>
</tr>
<tr>
<td>$e_1$</td>
<td>$\mathcal{R}_5$</td>
<td>B</td>
<td>C</td>
</tr>
<tr>
<td>$e_1$</td>
<td>$\mathcal{R}_2$</td>
<td>B</td>
<td>A</td>
</tr>
<tr>
<td>$e_1$</td>
<td>$\mathcal{R}_5$</td>
<td>B</td>
<td>A</td>
</tr>
<tr>
<td>$e_2$</td>
<td>$\mathcal{R}_2$</td>
<td>B</td>
<td>C</td>
</tr>
<tr>
<td>$e_2$</td>
<td>$\mathcal{R}_5$</td>
<td>B</td>
<td>C</td>
</tr>
<tr>
<td>$e_2$</td>
<td>$\mathcal{R}_5$</td>
<td>B</td>
<td>A</td>
</tr>
<tr>
<td>$e_2$</td>
<td>$\mathcal{R}_2$</td>
<td>B</td>
<td>A</td>
</tr>
</tbody>
</table>

Retrieval function based on binary ($\mathcal{R}_2$) and multi-valued relevance ($\mathcal{R}_5$) scales
<table>
<thead>
<tr>
<th>retr. fct.</th>
<th>relev. scale</th>
<th>learn. sample</th>
<th>test sample</th>
<th>$R_{\mu_{\text{norm}}}$</th>
<th>$R_{M_{\text{norm}}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$e_1$</td>
<td>$\mathcal{R}_2$</td>
<td>B/10</td>
<td>C</td>
<td>0.699</td>
<td>0.713</td>
</tr>
<tr>
<td>$e_1$</td>
<td>$\mathcal{R}_5$</td>
<td>B/10</td>
<td>C</td>
<td>0.695</td>
<td>0.718</td>
</tr>
<tr>
<td>$e_1$</td>
<td>$\mathcal{R}_2$</td>
<td>B/10</td>
<td>A</td>
<td>0.725</td>
<td>0.713</td>
</tr>
<tr>
<td>$e_1$</td>
<td>$\mathcal{R}_5$</td>
<td>B/10</td>
<td>A</td>
<td>0.730</td>
<td>0.732</td>
</tr>
<tr>
<td>$e_2$</td>
<td>$\mathcal{R}_2$</td>
<td>B/10</td>
<td>C</td>
<td>0.689</td>
<td>0.692</td>
</tr>
<tr>
<td>$e_2$</td>
<td>$\mathcal{R}_5$</td>
<td>B/10</td>
<td>C</td>
<td>0.702</td>
<td>0.689</td>
</tr>
<tr>
<td>$e_2$</td>
<td>$\mathcal{R}_5$</td>
<td>B/10</td>
<td>A</td>
<td>0.711</td>
<td>0.727</td>
</tr>
<tr>
<td>$e_2$</td>
<td>$\mathcal{R}_2$</td>
<td>B/10</td>
<td>A</td>
<td>0.696</td>
<td>0.716</td>
</tr>
</tbody>
</table>

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
6.5.2 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 $\vec{z}_k = (z_{k1}, \ldots, z_{kn}) = \begin{cases} 
1, & \text{if } t_i \in q_k^T \\
0, & \text{otherwise}
\end{cases}$

$d_m$ 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:
\[ P(R|\vec{z}_k, d_m) = P(R|d_m) \cdot \frac{P(\vec{z}_k | R, d_m)}{P(\vec{z}_k | d_m)} \]  

\( P(R|d_m) \) 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:

\[ P(\vec{z}_k | R, d_m) = \prod_{i=1}^{n} P(z_{k_i} | R, d_m) \]
\[ P(R|\vec{z}_k, d_m) = \frac{P(R|d_m)}{P(\vec{z}_k|d_m)} \cdot \prod_{i=1}^{n} P(z_{ki}|R, d_m) \]
\[ = \frac{P(R|d_m)}{P(\vec{z}_k|d_m)} \cdot \prod_{i=1}^{n} \frac{P(R|z_{ki}, d_m)}{P(R|d_m)} \cdot P(z_{ki}|d_m) \]

\[ P(\vec{z}_k|d_m) \text{ and } P(z_{ki}|d_m) \text{ are independent of a specific document} \]
\[ \text{(since we always regard all documents w.r.t. a query)}: \]
\[ P(R|\vec{z}_k, d_m) = \frac{\prod_{i=1}^{n} P(z_{ki})}{P(\vec{z}_k)} \cdot P(R|d_m) \cdot \prod_{i=1}^{n} \frac{P(R|z_{ki}, d_m)}{P(R|d_m)} \]
\[ = \frac{\prod_{i=1}^{n} P(z_{ki})}{P(\vec{z}_k)} \cdot P(R|d_m) \cdot \prod_{z_{ki}=1} P(R|z_{ki} = 1, d_m) \cdot \prod_{z_{ki}=0} \frac{P(R|z_{ki} = 0, d_m)}{P(R|d_m)} \quad (4) \]
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:

$$\prod_{i=1}^{n} \frac{P(z_{k_i})}{P(\vec{z}_k)} = c_k$$
\[ P(R|z_{k_i} = 1, d_m) = P(R|t_i, d_m): \]
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_{T_m} \text{ should contain at least those terms from } T \text{ 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 \( t_i \notin d_{T_m}^T \):
\[
P(R|q_k, d_m) = c_k \cdot P(R|d_m) \cdot \prod_{t_i \in q_k \cap d_m^T} \frac{P(R|t_i, d_m)}{P(R|d_m)}. \quad (5)
\]
application of BII model in this form nearly impossible, because of lack of relevance information for specific term-document pairs.
6.5.3 A description-oriented indexing approach

Darmstadt Indexing Approach

\[(t, d) \rightarrow P(C \mid t, d) \rightarrow P(C \mid x(t, d))\]

description \quad \rightarrow \quad decision

\[x(t, d)\]

relevance description

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 \( x \)

→ 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 $L \subset Q \times D \times R$
(query-document with relevance judgements)$L = \{(q_k, d_m, r_{km})\}$.

form relevance descriptions for the terms common to query and document:

bag of relevance descriptions with relevance judgements
$L^x = [(x(t_i, d_m), r_{km})|t_i \in q_k^T \cap d_m^T \land (q_k, d_m, r_{km}) \in L]$.

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

<table>
<thead>
<tr>
<th>query</th>
<th>doc.</th>
<th>judg.</th>
<th>term</th>
<th>$\vec{x}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$q_1$</td>
<td>$d_1$</td>
<td>$R$</td>
<td>$t_1$</td>
<td>(1, 1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$t_2$</td>
<td>(0, 1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$t_3$</td>
<td>(1, 2)</td>
</tr>
<tr>
<td>$q_1$</td>
<td>$d_2$</td>
<td>$\bar{R}$</td>
<td>$t_1$</td>
<td>(0, 2)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$t_3$</td>
<td>(1, 1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$t_4$</td>
<td>(0, 1)</td>
</tr>
<tr>
<td>$q_2$</td>
<td>$d_1$</td>
<td>$R$</td>
<td>$t_2$</td>
<td>(0, 2)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$t_5$</td>
<td>(0, 2)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$t_6$</td>
<td>(1, 1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$t_7$</td>
<td>(1, 2)</td>
</tr>
<tr>
<td>$q_2$</td>
<td>$d_3$</td>
<td>$\bar{R}$</td>
<td>$t_5$</td>
<td>(0, 1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$t_7$</td>
<td>(0, 1)</td>
</tr>
</tbody>
</table>

| $\vec{x}$ | $P(R|\vec{x})$ |
|------------|-----------------|
| $(0, 1)$   | $EX$ | $EBII$ |
| $(0, 2)$   | $1/4$ | $1/3$ |
| $(1, 1)$   | $2/3$ | $1/2$ |
| $(1, 2)$   | $2/3$ | $2/3$ |
| $(0, 1)$   | $1$    | $1$    |
2 different event spaces:

$E_{BII}$: equiprobable query-document pairs

$E_X$: equiprobable relevance descriptions
Indexing function

Data used (same as in SMART):

\[ t_{f_{mi}}: \text{ within-document frequency (wdf) of } t_i \text{ in } d_m. \]

\[ \max t_{f_{m}}: \text{ maximum wdf } t_{f_{mi}} \text{ of all terms } t_i \in d_m^T. \]

\[ n_i: \text{ number of documents in which } t_i \text{ occurs.} \]

\[ |D|: \text{ number of documents in the collection.} \]

\[ |d_m^T|: \text{ number of different terms in } d_m. \]

\[ t_{a_{mi}}: = 1, \text{ if } t_i \text{ occurs in the title of } d_m, \text{ and } 0 \text{ otherwise.} \]
Relevance description elements:

\[ x_1 = tf_{mi} \]
\[ x_2 = 1/\max tf_m \]
\[ x_3 = \log(n_i/|D|) \]
\[ x_4 = \log |d^T_m| \]
\[ x_5 = ta_{mi} \]

Indexing functions:

\[ e_L = a_0 + a_1 tf_{mi} + a_2/\max tf_m + a_3 \log(n_i/|D|) + a_4 \log |d^T_m|, \]
\[ e_{ta} = a_0 + a_1 tf_{mi} + a_2/\max tf_m + a_3 \log(n_i/|D|) + a_4 \log |d^T_m| + a_5 ta_{mi} \]
Retrieval functions

$q^T_k$ — set of query terms

d$^T_m$ — set of document terms

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

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

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

$$q(q_k, d_m) = \sum_{t_i \in q^T_k \cap d^T_m} c_{ki} \cdot u_{mi}$$

- $q_{bin}$ binary query term weights ($c_{ki} = 1$ for all $t_i \in q^T_k$)
- $q_{tf}$ $c_{ki} =$ within-query frequency of $t_i$ in $q_k$
- $q_{tfidf}$ $c_{ki} = tfidf$ with within-query frequencies
## Experimental results

<table>
<thead>
<tr>
<th>collection</th>
<th>$tf \times idf$</th>
<th>$e_L$</th>
<th>$e_{ta}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$Q_{tfidf}$</td>
<td>$Q_{bin}$</td>
<td>$Q_{tf}$</td>
</tr>
<tr>
<td>CACM</td>
<td>0.2963</td>
<td>0.3046</td>
<td>0.3371</td>
</tr>
<tr>
<td></td>
<td>+ 2.8%</td>
<td>+ 13.8%</td>
<td>+ 10.8%</td>
</tr>
<tr>
<td>CISI</td>
<td>0.2099</td>
<td>0.1731</td>
<td>0.2288</td>
</tr>
<tr>
<td></td>
<td>- 17.5%</td>
<td>+ 9.0%</td>
<td>- 2.2%</td>
</tr>
<tr>
<td>CRAN</td>
<td>0.3816</td>
<td>0.4265</td>
<td>0.4293</td>
</tr>
<tr>
<td></td>
<td>+ 11.8%</td>
<td>+ 12.5%</td>
<td>+ 2.8%</td>
</tr>
<tr>
<td>INSPEC</td>
<td>0.2489</td>
<td>0.2307</td>
<td>0.2708</td>
</tr>
<tr>
<td></td>
<td>- 7.3%</td>
<td>+ 8.8%</td>
<td>+ 0.5%</td>
</tr>
<tr>
<td>NPL</td>
<td>0.2138</td>
<td>0.2834</td>
<td>0.2834</td>
</tr>
<tr>
<td></td>
<td>+ 32.6%</td>
<td>+ 32.6%</td>
<td>+ 11.4%</td>
</tr>
</tbody>
</table>
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 $\rho_{tf}$ better than SMART approach
- SMART approach works without relevance information
- probabilistic indexing can be extended easily to more complex representations
6.5.4 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
→ extended event space $Q \times D \times I$:

$Q$ set of queries
$D$ 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 \( q_k^J = \{(d_m, r(d_m, q_k))\} \)

- 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.
\[ O(R|q_k, \bar{x}) = \]

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

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

\[ u_{m_i} = P(C_i|x_i=d_{m_i}): \text{probabilistic indexing weight of } t_i \text{ w.r.t. } d_m \]

\[ = \text{probability that arbitrary indexer assigned } t_i \text{ to } d_m \]

\[ q_i = P(C_i): \text{probability that arbitrary indexer assigned } t_i \text{ to arbitrary document} \]

\[ = \text{average indexing weight of } t_i \text{ in the collection} \]

\[ p_{ik} = P(C_i|R, q_k): \text{probability that arbitrary indexer assigned } t_i \text{ to arbitrary document relevant to } q_k \]

\[ = \text{average indexing weight of } t_i \text{ in relevant documents} \]

\[ r_{ik} = P(C_i|\bar{R}, q_k): \text{probability that arbitrary indexer assigned } t_i \text{ to arbitrary doc. nonrelevant to } q_k \]

\[ = \text{average indexing weight of } t_i \text{ in nonrelevant docs} \]

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

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

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

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

\[ D^R_k \] — set of docs. judged relevant w.r.t. \( q_k \)

\[
q_i = \frac{1}{|D|} \sum_{d_j \in D} u_{ji}
\]

\[
p_{ik} = \frac{1}{|D^R_k|} \sum_{d_j \in D^R_k} u_{ji}
\]
6.6 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 $\rightarrow$ parameters)
**Estimation methods**

collection of documents with features $e_i$

estimation of $P(e_i | e_j)$

*BIR model:*

$e_j = \text{relevance} / \text{non-relevance} \ w.r.t. \ q_k$

$e_i = \text{presence} / \text{absence} \ of \ a \ term \ t_i$

$g$ 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:

\[
g(p\mid h, f) = \frac{p^{a-1}(1 - p)^{b-1}(f^h) p^{h}(1 - p)^{f-h}}{\int_0^1 p^{a-1}(1 - p)^{(b-1)}(f^h) p^{h}(1 - p)^{f-h} dp}
\]

\[
B(a, b) = \int_0^1 p^{a-1}(1 - p)^{b-1} dp \text{ yields}
\]

posterior distribution:

\[
g(p\mid h, f) = \frac{p^{h+a-1}(1 - p)^{f-h+b-1}}{B(h + a, f - h + b)}
\]
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:

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

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

<table>
<thead>
<tr>
<th>$f$</th>
<th>$Z_g'(h, f)$</th>
<th>$a$</th>
<th>$b$</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>139</td>
<td>-3.42</td>
<td>0.47</td>
</tr>
<tr>
<td>5</td>
<td>341</td>
<td>2.64</td>
<td>5.85</td>
</tr>
<tr>
<td>6</td>
<td>349</td>
<td>-1.54</td>
<td>1.48</td>
</tr>
<tr>
<td>7</td>
<td>546</td>
<td>-1.61</td>
<td>1.33</td>
</tr>
<tr>
<td>8</td>
<td>526</td>
<td>-1.54</td>
<td>1.76</td>
</tr>
<tr>
<td>9</td>
<td>816</td>
<td>-1.29</td>
<td>1.84</td>
</tr>
<tr>
<td>10</td>
<td>374</td>
<td>-1.46</td>
<td>1.31</td>
</tr>
</tbody>
</table>

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 ($g = 78 000$).
<table>
<thead>
<tr>
<th>sample</th>
<th>$Z_{g'}$</th>
<th>estimate</th>
<th>$s^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X$</td>
<td>2852</td>
<td>$p_{ML}$</td>
<td>0.271</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$p_{L_1}$</td>
<td>0.231</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$p_{opt}$</td>
<td>0.231</td>
</tr>
<tr>
<td>$Z$</td>
<td>2709</td>
<td>$p_{ML}$</td>
<td>0.265</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$p_{L_1}$</td>
<td>0.226</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$p_{opt}$</td>
<td>0.224</td>
</tr>
</tbody>
</table>

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